# BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

# IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.**

*Proc. Natl. Acad. Sci. USA*
Vol. 94, pp. 8945–8947, August 1997
Applied Biological Sciences

# Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR

DEVAL A. LASHKARI*†, JOHN H. MCCUSKER‡, AND RONALD W. DAVIS*§

*Departments of Genetics and Biochemistry, Beckman Center, Stanford University, Stanford, CA 94305; and ‡Department of Microbiology, 3020 Duke University Medical Center, Durham, NC 27710

*Contributed by Ronald W. Davis, May 20, 1997*

**ABSTRACT** **The recent ability to sequence whole genomes allows ready access to all genetic material. The approaches outlined here allow automated analysis of sequence for the synthesis of optimal primers in an automated multiplex oligonucleotide synthesizer (AMOS). The efficiency is such that all ORFs for an organism can be amplified by PCR. The resulting amplicons can be used directly in the construction of DNA arrays or can be cloned for a large variety of functional analyses. These tools allow a replacement of single-gene analysis with a highly efficient whole-genome analysis.**

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of *Saccharomyces cerevisiae*, *Escherichia coli*, *Haemophilus influenzae* (1), *Mycoplasma genitalium* (2), and *Methanococcus jannaschii* (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well, including the nematode *Caenorhabditis elegans* (4) and the small flowering plant *Arabidopsis thaliana* (5). This massive and increasing amount of sequence information allows the development of novel experimental approaches to identify gene function.

One standard use of genome sequence data is to attempt to identify the functions of predicted open reading frames (ORFs) within the genome by comparison to genes of known function. Such a comparative analysis of all ORFs to existing sequence data is fast, simple, and requires no experimentation and is therefore a reasonable first step. While finding sequence homologies/motifs is not a substitute for experimentation, noting the presence of sequence homology and/or sequence motifs can be a useful first step in finding interesting genes, in designing experiments and, in some cases, predicting function. However, this type of analysis is frequently uninformative. For example, over one-half of new ORFs in *S. cerevisiae* have no known function (6). If this is the case in a well studied organism such as yeast, the problem will be even worse in organisms that are less well studied or less manipulable. A large, experimentally determined gene function database would make homology/motif searches much more useful.

Experimental analysis must be performed to thoroughly understand the biological function of a gene product. Scaling up from classical "cottage industry" one-gene-oriented approaches to whole-genome analysis would be very expensive and laborious. It is clear that novel strategies are necessary to efficiently pursue the next phase of the genome projects—whole-genome experimental analysis to explore gene expression, gene product function, and other genome functions. Model organisms, such as *S. cerevisiae*, will be extremely important in the development of novel whole-genome analysis techniques and, subsequently, in improving our understanding of other more complex and less manipulable organisms.

The genome sequence can be systematically used as a tool to understand ORFs, gene product function, and other genome regions. Toward this end, a directed strategy has been developed for exploiting sequence information as a means of providing information about biological function (Fig. 1). Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons—they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay (7). As a pilot study, synthetic primers were made on the 96-well automated multiplex oligonucleotide synthesizer (AMOS) instrument (8) (Fig. 2). These oligonucleotides were used to amplify each ORF on yeast chromosome V. The current version of this instrument can synthesize three plates of 96 oligonucleotides each (25 bases) in an 8-hr day. The amplification of the entire set of PCR products was then analyzed by gel electrophoresis (Fig. 3). Successful amplification of the proper length product on the first attempt was 95%. This project demonstrates that one can go directly from sequence information to biological analysis in a truly automated, totally directed manner.

These amplicons can be incorporated directly in arrays or the amplicons can be cloned. If the amplicons are to be cloned, novel sequences can be incorporated at the 5′ end of the oligonucleotide to facilitate cloning. One potential problem with cloning PCR products is that the cloned amplicons may contain sequence alterations that diminish their utility. One option would be to resequence each individual amplicon. However, this is expensive, inefficient, and time consuming. A faster, more cost-effective, and more accurate approach is to apply comparative sequencing by denaturing HPLC (9). This method is capable of detecting a single base change in a 2-kb heteroduplex. Longer amplicons can be analyzed by use of appropriate restriction fragments. If any change is detected in a clone, an alternate clone of the same region can be analyzed. Modifying the system to allow high throughput analysis by denaturing HPLC is also relatively simple and straightforward.

If amplicons are used directly on arrays without cloning, it is important to note that, even if single PCR product bands are observed on gels, the PCR products will be contaminated with various amounts of other sequences. This contamination has the potential to affect the results in, for example, expression

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.



†Present address: Synteni, Inc., 6519 Dumbarton Circle, Fremont, CA 94555.
§To whom reprint requests should be addressed at: Department of Biochemistry, Beckman Center, B400, Stanford University, Stanford, CA 94305-5307. e-mail: gilbert@cmgm.stanford.edu.

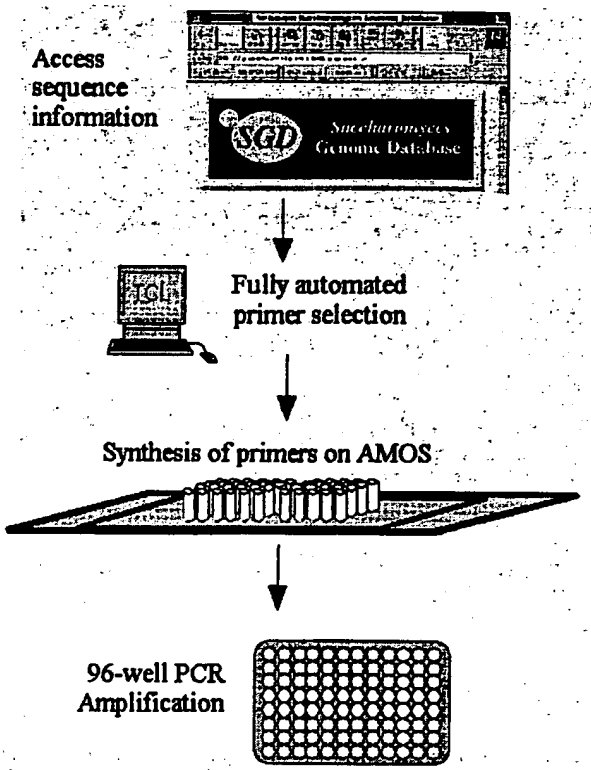


FIG. 1. Overview of systematic method for isolating individual genes. Sequence information is obtained automatically from sequence databases. The data are input into primer selection software specifically designed to target ORFs as designated by database annotations. The output file containing the primer information is directly read by a high-throughput oligonucleotide synthesizer, which makes the oligonucleotides in 96-well plates (AMOS, automated multiplex oligonucleotide synthesizer). The forward and reverse primers are synthesized in the same location on separate plates to facilitate the downstream handling of primers. The amplicons are generated by PCR in 96-well plates as well.

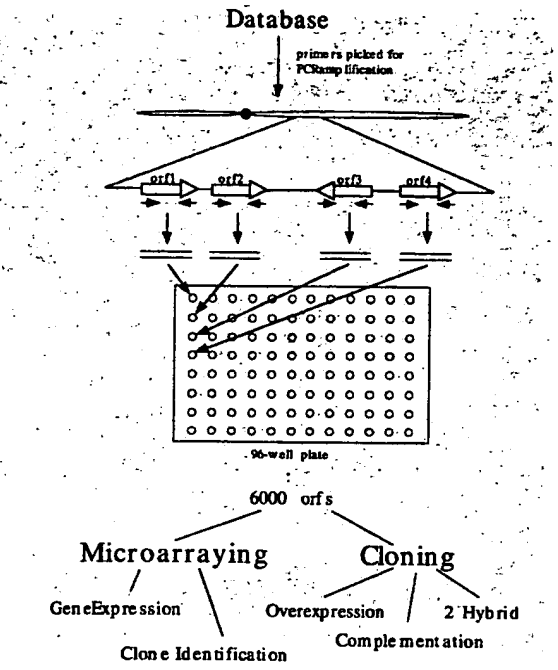


FIG. 2. Overall approach for using database of a genome to direct biological analysis. The synthesis of the 6,000 ORFs (orfs) for each gene of *S. cerevisiae* can be used in many applications utilizing both cloning and microarraying technology.

analysis. On the other hand, direct use of the amplicons is much less labor intensive and greatly decreases the occurrence of mistakes in clone identification, a ubiquitous problem associated with large clone set archiving and retrieving.

Any large-scale effort to capture each ORF within a genome must rely on automation if cost is to be minimized while efficiency is maximized. Toward that end, primers targeting ORFs were designed automatically using simple new scripts and existing primer selection software. These script-selected primer sequences were directly read by the high-throughput synthesizer and the forward and reverse primers were synthesized in separate plates in corresponding wells to facilitate automated pipetting and PCR amplifications. Each of the resulting PCR products, generated with minimum labor, contains a known, unique ORF.

Large-scale genome analysis projects are dependent on newly emerging technologies to make the studies practical and economically feasible. For example, the cost of the primers, a significant issue in the past, has been reduced dramatically to make feasible this and other projects that require tens of thousands of oligonucleotides. Other methods of high-throughput analysis are also vital to the success of functional analysis projects, such as microarraying and oligonucleotide chip methods (10–14).

Changes in attitude are also required. One of the major costs of commercial oligonucleotides is extensive quality control such that virtually 100% of the supplied oligonucleotides are successfully synthesized and work for their intended purpose. Considerable cost reduction can be obtained by simply decreasing the expected successful synthesis rate to 95–97%. One can then achieve faster and cheaper whole genome coverage by simply adding a single quality control at the end of the experiment and batching the failures for resynthesis.

The directed nature of the amplicon approach is of clear advantage. The sequence of each ORF is analyzed automatically, and unique specific primers are made to target each ORF. Thus, there is relatively little time or labor involved—for example, no random cloning and subsequent screening is required because each product is known. In the test system, primers for 240 ORFs from chromosome V were systematically synthesized, beginning from the left arm and continuing through to the right arm. At no point was there any manual analysis of sequence information to generate the collection. In many ways, now that the sequence is known, there is no need for the researcher to examine it.

These amplicons can be arrayed and expression analysis can be done on all arrayed ORFs with a single hybridization (10). Those ORFs that display significant differential expression patterns under a given selection are easily identified without the laborious task of searching for and then sequencing a clone. Once scaled up, the procedure provides even greater returns on effort, because a single hybridization will ultimately provide a "snapshot" of the expression of all genes in the yeast genome. Thus, the limiting factor in whole genome analysis will not be the analysis process itself, but will instead be the ability of researchers to design and carry out experimental selections.

Current expression and genetic analysis technologies are geared toward the analysis of single genes and are ill suited to analyze numerous genes under many conditions. Additional difficulties with current technologies include: the effort and expense required to analyze expression and make mutants, the potential duplication of effort if done by different laboratories, and the possibility of conflicting results obtained from different laboratories. In contrast, whole genome analysis not only is more efficient, it also provides data of much higher quality; all genes are assayed and compared in parallel under exactly

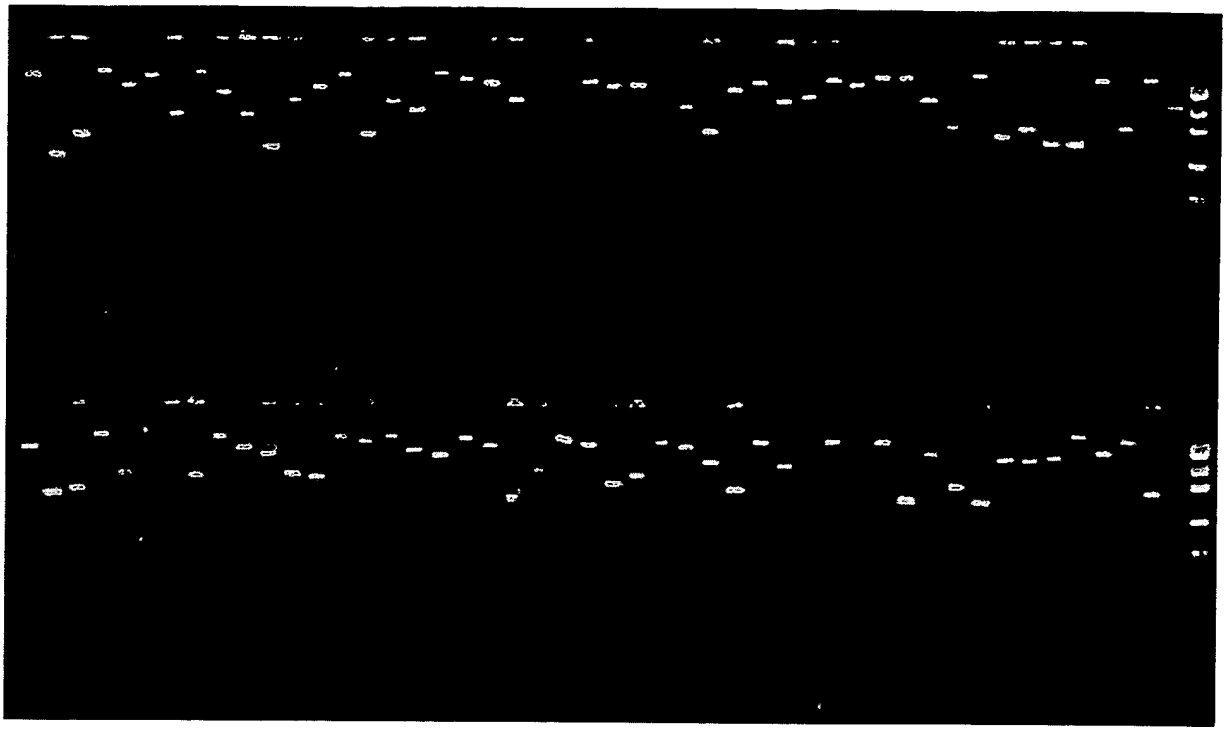

FIG. 3. Gel image of amplifications. Using the method described in Fig. 1, amplicons were generated for ORFs of *S. cerevisiae* chromosome V. One plate of 96 amplification reactions is shown.

the same conditions. In addition, amplicons have many applications beyond gene expression. For example, one recent approach is to incorporate a unique DNA sequence tag, synthesized as part of each gene specific primer, during amplification. The tags or molecular bar codes, when reintroduced into the organism as a gene deletion or as a gene clone, can be used much more efficiently than individual mutations or clones because pools of tagged mutants or transformants can be analyzed in parallel. This parallel analysis is possible because the tags are readily and quantitatively amplified even in complex mixtures of tags (13).

These ORF genome arrays and oligonucleotide tagged libraries can be used for many applications. Any conventional selection applied to a library that gives discrete or multiple products can use these technologies for a simple direct readout. These include screens and selections for mutant complementation, overexpression suppression (15, 16), second-site suppressors, synthetic lethality, drug target overexpression (17), two-hybrid screens (18), genome mismatch scanning (19), or recombination mapping.

The genome projects have provided researchers with a vast amount of information. These data must be used efficiently and systematically to gain a truly comprehensive understanding of gene function and, more broadly, of the entire genome which can then be applied to other organisms. Such global approaches are essential if we are to gain an understanding of the living cell. This understanding should come from the viewpoint of the integration of complex regulatory networks, the individual roles and interactions of thousands of functional gene products, and the effect of environmental changes on both gene regulatory networks and the roles of all gene products. The time has come to switch from the analysis of a single gene to the analysis of the whole genome.

Support was provided by National Institutes of Health Grants R37H60198 and P01H600205.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **269,** 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* **270,** 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* **273,** 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. & Waterston, R. (1992) *Nature (London)* **356,** 37–41.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* **106,** 1241–1255.
6. Oliver, S. (1996) *Nature (London)* **379,** 597–600.
7. Lashkari, D. A. (1996) Ph.D. dissertation (Stanford Univ., Stanford, CA).
8. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W. & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 7912–7915.
9. Oefner, P. J. & Underhill, P. A. (1995) *Am. J. Hum. Genet.* **57,** A266.
10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270,** 467–470.
11. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Science* **251,** 767–773.
12. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. (1996) *Science* **274,** 610–614.
13. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. (1996) *Nat. Genet.* **14,** 450–456.
14. Smith, V., Chou, K., Lashkari, D., Botstein, D. & Brown, P. O. (1996) *Science* **274,** 2069–2074.
15. Magdolen, V., Drubin, D. G., Mages, G. & Bandlow, W. (1993) *FEBS Lett.* **316,** 41–47.
16. Ramer, S. W., Elledge, S. J. & Davis, R. W. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 11589–11593.
17. Rine, J., Hansen, W., Hardeman, E. & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 6750–6754.
18. Fields, S. & Song, O. (1989) *Nature (London)* **340,** 245–246.
19. Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P. & Brown, P. O. (1994) *Nat. Genet.* **4,** 11–18.

MOLECULAR CARCINOGENESIS 24:153–159 (1999)

IN PERSPECTIVE
Claudio J. Conti, Editor

# Microarrays and Toxicology: The Advent of Toxicogenomics

**Emile F. Nuwaysir,[1] Michael Bittner,[2] Jeffrey Trent,[2] J. Carl Barrett,[1] and Cynthia A. Afshari[1]**

[1]*Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina*
[2]*Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland*

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog. 24:153–159, 1999.* 

Key words: toxicology; gene expression; animal bioassay

## INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae,* was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cervisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNAse protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10–12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

## MICROARRAY DEVELOPMENT AND APPLICATIONS

### cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3′ regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease–related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cervisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

### Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnostics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)+ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and BRCA1 [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

## THE USE OF MICROARRAYS IN TOXICOLOGY

### Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on relative fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

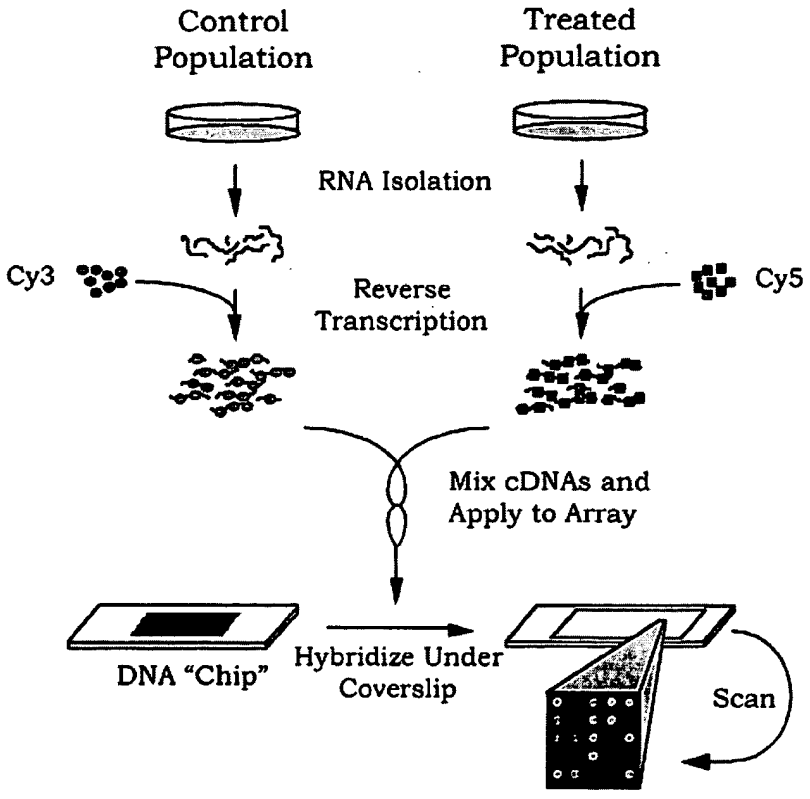


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illustrative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

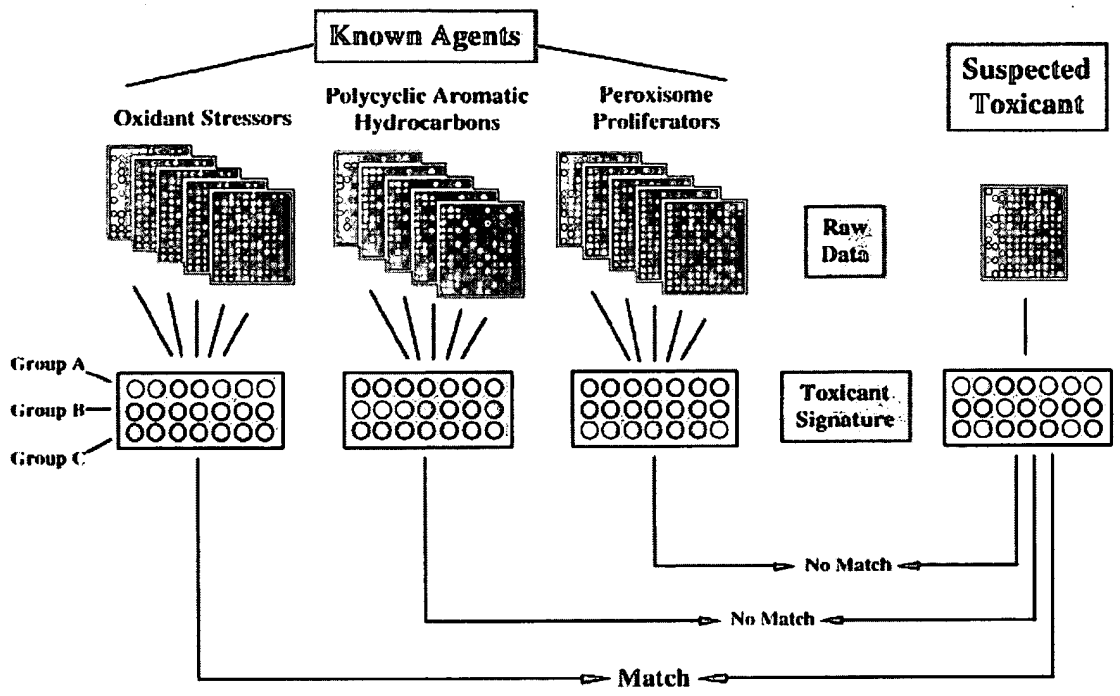


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

## Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

| Gene category | No. of genes on chip |
|---|---|
| Apoptosis | 72 |
| DNA replication and repair | 99 |
| Oxidative stress/redox homeostasis | 90 |
| Peroxisome proliferator responsive | 22 |
| Dioxin/PAH responsive | 12 |
| Estrogen responsive | 63 |
| Housekeeping | 84 |
| Oncogenes and tumor suppressor genes | 76 |
| Cell-cycle control | 51 |
| Transcription factors | 131 |
| Kinases | 276 |
| Phosphatases | 88 |
| Heat-shock proteins | 23 |
| Receptors | 349 |
| Cytochrome P450s | 30 |

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive in vivo test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia,* and *Arabadopsis* could also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

### Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

### Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

## FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

## ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

## REFERENCES

1. http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html
2. http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996;274:546, 563–567.
5. http://www.perkin-elmer.com/press/prc5448.html
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257:967–971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 1996;6:492–503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. Gene 1995;156:207–213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. Science 1995;270:467–470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol 1997;15:1359–1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. Nat Biotechnol 1998;16:27–31.
14. http://www.synteni.com
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6:639–645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomedical Optics 1997;2:364–374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998;58:5009–5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996; 93:10614–10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 1997;94:2150–2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 1996;37:29–40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the E. coli genome. Genomics 1996;37:77–86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. Biotechniques 1994;17:328–329, 332–336.
25. http://www.resgen.com/
26. http://www.genomesystems.com/
27. http://www.clontech.com/
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022–5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767–773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci USA 1996;93:13555–13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 1995;19:442–447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.
34. http://www.mdyn.com/
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. Genomics 1996;33:445–456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. Nat Genet 1998;18:155–158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 1996;7:244–255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 1996;14:441–447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med 1996;2:753–759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. Science 1998;281:1194–1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. Environ Health Perspect 1990;86:313–321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19–23.
45. http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 1996;382:722–725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. J Natl Cancer Inst 1993;85:1159–1164.
48. http://www.niehs.nih.gov/envgenom/home.html

ELSEVIER

Toxicology Letters 112–113 (2000) 467–471

Toxicology Letters

www.elsevier.com/locate/toxlet

# Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

*Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA*

**Abstract**

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers.


*Keywords:* Proteomics; Genomics; Toxicology

## 1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

* Corresponding author. Tel.: +1-301-4245989; fax: +1-301-7624892.
*E-mail address:* steiner@lsbc.com (S. Steiner)

0378-4274/00/$ - see front matter 
PII: S0378-4274(99)00236-2

## 2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

## 3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

## 4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

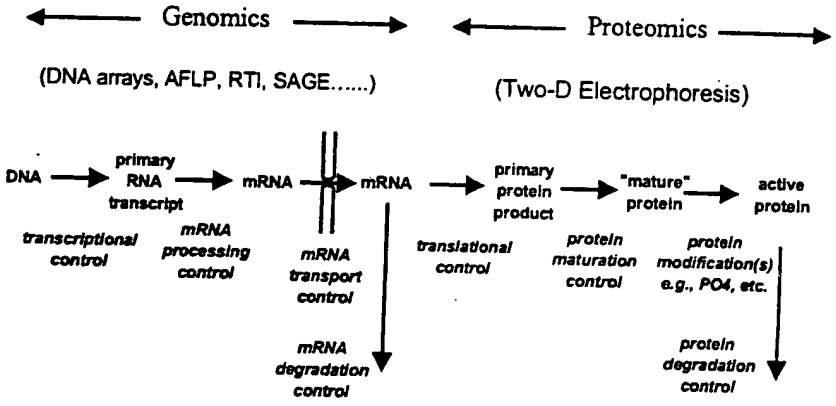


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

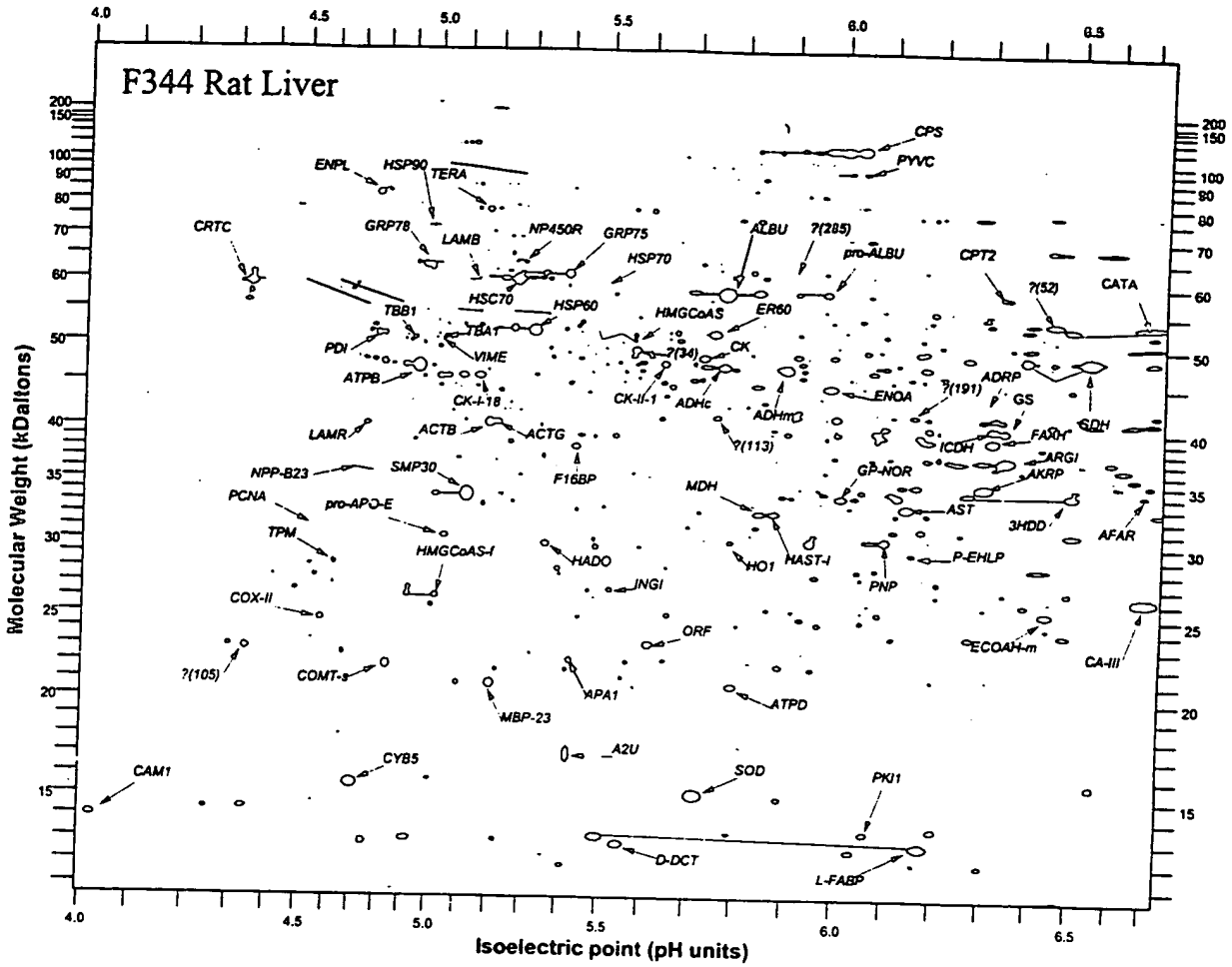


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

## 5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplifi-

cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

## 6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug development, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

## 7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trails.

## References

Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. Electrophoresis 19, 1998–2003.

Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18, 533–537.

Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. Electrophoresis 12, 907–930.

Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. J. Cell. Biochem. Suppl. 22, 108–116.

Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. Toxicol. Appl. Pharmacol. 137, 75–89.

Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. Life Sci. 63, 2243–2250.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. Science 274, 610–614.

Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. Electrophoresis 19, 355–363.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251, 767–773.

Mann, M., Hojrup, P., Roepsdorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338–345.

Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. Electrophoresis 14, 157–161.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expresssion patterns with a complementary DNA microarray. Science 251, 467–470.

Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 6, 639–645.

Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Rayrnackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. Biochem. Biophys. Res. Commun. 218, 777–782.

Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. Biochem. Pharmacol. 51, 253–258.

Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. Curr. Biol. 6, 1543–1544.

Docket No.: PF-0594 USN
USSN: 09/786,797
Ref. No. 5 of 9

# Application of DNA Arrays to Toxicology

*John C. Rockett and David J. Dix*

Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations. *Key words:* DNA arrays, gene arrays, microarrays, toxicology. *Environ Health Perspect* 107:681–685 (1999). [Online 6 July 1999]
*http://ehpnet1.niehs.nih.gov/docs/1999/107p681-685rockett/abstract.html*

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The potential importance of arrays is enormous and has been highlighted by the recent publication of an entire *Nature Genetics* supplement dedicated to the technology (*1*). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the high ratio of review and press articles to actual data papers. Even so, the potential they offer has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained in the workshop presentations should provide aid and insight into arrays in general and their application to toxicology in particular.

## Array Elements

In the context of molecular biology, the word "array" is normally used to refer to a series of DNA or protein elements firmly attached in a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (*2*). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

## Array Printing

Arrays are typically printed on one of two types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

Address correspondence to J. Rockett, Reproductive Toxicology Division (MD-72), National Health and Environmental Effects Research Laboratory, U.S. EPA, Research Triangle Park, NC 27711 USA. Telephone: (919) 541-2678. Fax: (919) 541-4017. E-mail: rockett.john@epa.gov

The authors thank R. Kavlock for envisioning the application of array technology to toxicology at the U.S. Environmental Protection Agency. We also thank T. Wall and B. Deitz for administrative assistance.

This document has been reviewed in accordance with EPA policy and approved for publication. Mention of companies, trade names, or products does not signify endorsement of such by the EPA.

Received 23 March 1999; accepted 22 April 1999.

coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic MicroSystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrayers, as described by Brown (*3*). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of > 2,500 spots/cm$^2$ may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental arrays; the first 100 or so spots of a new run tend to be somewhat variable. Factors effecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

## Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., $^{33}$P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosporimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptenylated, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (*4*). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of polyA$^+$ RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

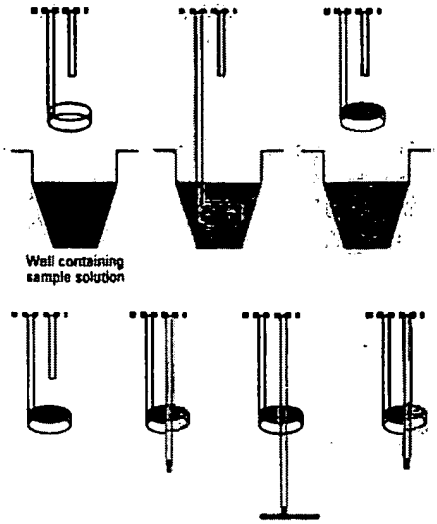


**Figure 1.** Genetic Microsystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution, with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (*14*), with permission from Genetic Microsystems.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reactions may also be used. Fluorophore- or radiolabeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5′ end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case fluor-labeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy)3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series of labeled nucleotides with a wide range of excitation and emission spectra which may prove to function as well as the Cy dyes.

## Analysis of DNA Microarrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Because a typical spot on a microarray can contain > $10^8$ molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of < 1 fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

## Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the entire genome of the species they are being used to study. For this reason, they have strong support among researchers utilizing yeast and *Caenorhabditis elegans* (*5*). The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (*6,7*). With both of these species, it is relatively easy to perturb individual gene expression. Indeed, *C. elegans* knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

Table 1. Advantages and disadvantages of different microarray scanning systems.

| | | Nonconfocal laser scanner | |
|---|---|---|---|
| Advantages | Few moving parts | Relatively simple optics | Small depth of focus reduces artifacts |
| | Fast scanning of bright samples | | May have high light collection efficiency |
| Disadvantages | Less appropriate for dim samples | Low light collection efficiency | Small depth of focus requires scanning precision |
| | Optical scatter can limit performance | Background artifacts not rejected | |
| | | Resolution typically low | |

CCD, charge-coupled device.
From Kawasaki (*13*).

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomenclature is often confusing in that single genes are allocated multiple names (usually as a result of independent discovery by different laboratories), and there was a call for standardization of gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be transferred onto arrays and used to screen any chosen system. The EPA MicroArray Consortium (EPAMAC) is assembling testes transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (*8*) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially mollified this deficiency. ESTs are cDNAs expressed in a given tissue that, although they may share some degree of sequence similarity to previously characterized genes, have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This can enable the identification of previously uncharacterized genes that may have biologic

significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

- Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.
- Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, noncontaminated, sequence-verified, species/cell/tissue/field-specific clones.
- Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.
- Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.
- Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.
- Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a nonrelated species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).
- Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings. Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.
- Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.
- Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.
- Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

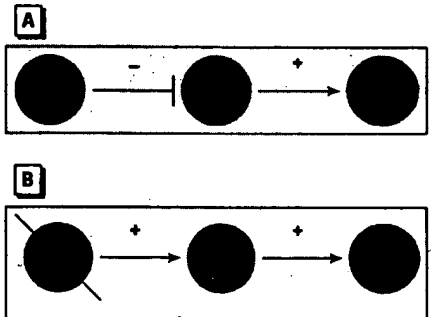


**Figure 2.** Potential effects of gene knockout within positively and negatively regulated gene expression networks. $i_1$ is limiting in wild type for expression of $i_2$. (*A*) A simple, two-component, linear regulatory network operating on gene $i_2$, where $i_1$ is a positive effector of $i_2$ and $j_n$ is either a positive or negative effector of $i_1$. This network could be deduced by examining the consequence of (*B*) deleting $j_n$ on the expression of $i_1$ and $i_2$, where the expression of $i_2$ would be decreased or increased depending on whether $j_n$ was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (*15*).

## EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or

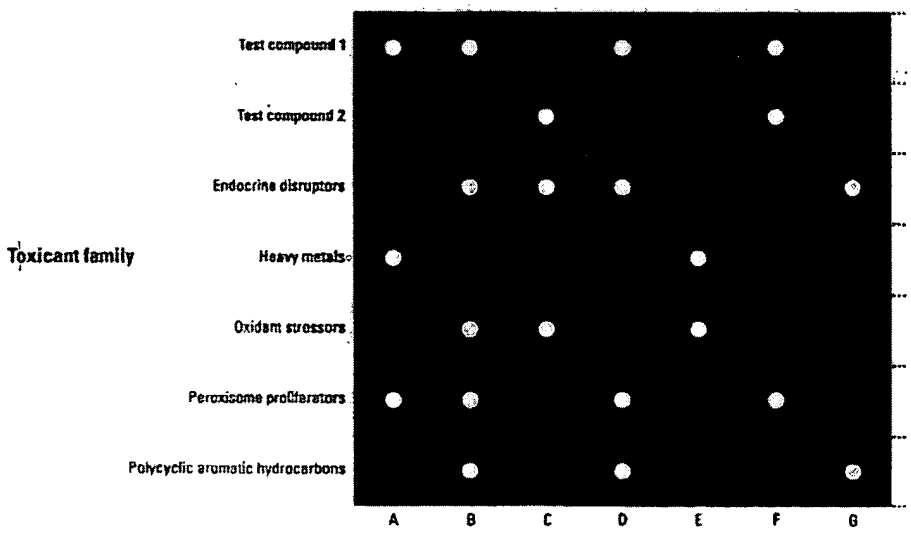


**Figure 3.** Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results, test compound 2 would be retained for further testing and test compound 1 would be eliminated.

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (*9*), and how this may compromise sperm counts and quality following sexual maturation (*10*). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (*11*) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

## Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene *X* is related to the expression of gene *Y*, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals, and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach is less expensive than laying out capital for one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (*12*):

> Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

### Speakers

Cindy Afshari, NIEHS
Linda Birnbaum, U.S. EPA
Ron Butow, University of Texas Southwestern Medical Center
Alex Chenchik, Clontech Laboratories, Inc.
David Dix, U.S. EPA
Abdel Elkahloun, Research Genetics, Inc.
Sue Fenton, U.S. EPA
Norman Hecht, University of Pennsylvania
Pat Hurban, Paradigm Genetics, Inc.
Bob Kavlock, U.S. EPA
Ernie Kawasaki, General Scanning, Inc.
Steve Krawetz, Wayne State University
Nick Mace, Genetic Microsystems, Inc.
Scott Mordecai, Affymetrix, Inc.
Kevin Morgan, Glaxo Wellcome, Inc.
Elaine Poplin, Research Genetics, Inc.
Don Rose, Cartesian Technologies, Inc.
Jim Samet, U.S. EPA
Sam Ward, University of Arizona
Jeff Welch, U.S. EPA
Reen Wu, University of California at Davis
Tim Zacharewski, Michigan State University

### References and Notes


1. The chipping forecast. Nat Genet 21(Suppl 1):3–60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: www.ncbi.nlm.nih.gov/Schuler/UniGene [cited 22 March 1999].
3. Brown PO. The Brown Lab. Available: http://cmgm.Stanford.edu/pbrown [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics 51:313–324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: www.mcb.arizona.edu/wardlab/microarray.html [cited 22 March 1999].
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301 (1998).
7. Brown PO. The Full Yeast Genome on a Chip. Available: http://cmgm.stanford.edu/pbrown/yeastchip.html [cited 22 March 1999].
8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24(3):153–159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. Bioessays 20:555–561 (1998).
10. Zacharewski TR. Timothy R. Zacharewski. Available: www.bch.msu.edu/faculty/zachar.htm [cited 22 March 1999].
11. Kramer JA, Krawetz SA. RNA in spermatozoa: implications for the alternative haploid genome. Mol Hum Reprod 3:473–478 (1997).
12. Stipp D. Gene chip breakthrough. Fortune, March 31:56–73 (1997).
13. Kawasaki E (General Scanning Instruments, Inc., Watertown, MA). Unpublished data.
14. Flowers P, Overbeck J, Mace ML Jr, Pagliughi FM, Eggers WJE, Yonkers H, Honkanen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: http://www.geneticmicro.com/resources/html/coldspring.html [cited 22 March 1999].
15. Butow R (University of Texas Medical Center, Dallas, TX). Unpublished data.

**Subject: RE: [Fwd: Toxicology Chip]**
**Date:** Mon, 3 Jul 2000 08:09:45 -0400
**From:** "Afshari.Cynthia" <afshari@niehs.nih.gov>
**To:** "'Diana Hamlet-Cox'" <dianahc@incyte.com>

You can see the list of clones that we have on our 12K chip at
http://manuel.niehs.nih.gov/maps/guest/clonesrch.cfm
We selected a subset of genes (2000K) that we believed critical to tox response and basic cellular processes and added a set of clones and ESTs to this. We have included a set of control genes (80-) that were selected by the NHGRI because they did not change across a large set of array experiments. However, we have found that some of these genes change significantly after tox treatments and are in the process of looking at the variation of each of these 80+ genes across our experiments.
Our chips are constantly changing and being updated and we hope that our data will lead us to what the toxchip should really be.
I hope this answers your question.
Cindy Afshari

> ----------
> *From: Diana Hamlet-Cox*
> *Sent: Monday, June 26, 2000 8:52 PM*
> *To: afshari@niehs.nih.gov*
> *Subject: [Fwd: Toxicology Chip]*
>
> *Dear Dr. Afshari,*
>
> *Since I have not yet had a response from Bill Grigg, perhaps he was not the right person to contact.*
>
> *Can you help me in this matter? I don't need to know the sequences, necessarily, but I would like very much to know what types of sequences are being used, e.g., GPCRs (more specific?), ion channels, etc.*
>
> *Diana Hamlet-Cox*
>
> *-------- Original Message --------*
> *Subject: Toxicology Chip*
> *Date: Mon, 19 Jun 2000 18:31:48 -0700*
> *From: Diana Hamlet-Cox <dianahc@incyte.com>*
> *Organization: Incyte Pharmaceuticals*
> *To: grigg@niehs.nih.gov*
>
> *Dear Colleague:*
>
> *I am doing literature research on the use of expressed genes as pharmacotoxicology markers, and found the Press Release dated February 29, 2000 regarding the work of the NIEHS in this area. I would like to know if there is a resource I can access (or you could provide?) that would give me a list of the 12,000 genes that are on your Human ToxChip Microarray. In particular, I am interested in the criteria used to select sequences for the ToxChip, including any control sequences included in the microarray.*
>
> *Thank you for your assistance in this request.*
>
> *Diana Hamlet-Cox, Ph.D.*
> *Incyte Genomics, Inc.*
>
> --
>
> ==========

> this email message is for the sole use of the intended recipient s and
> may contain confidential and privileged information subject to
> attorney-client privilege. Any unauthorized review, use, disclosure or
> distribution is prohibited. If you are not the intended recipient,
> please contact the sender by reply email and destroy all copies of the
> original message.
> ==========
>
>

# Proteomics: a major new technology for the drug discovery process

Martin J. Page, Bob Amess, Christian Rohlff, Colin Stubberfield and Raj Parekh

Proteomics is a new enabling technology that is being integrated into the drug discovery process. This will facilitate the systematic analysis of proteins across any biological system or disease, forwarding new targets and information on mode of action, toxicology and surrogate markers. Proteomics is highly complementary to genomic approaches in the drug discovery process and, for the first time, offers scientists the ability to integrate information from the genome, expressed mRNAs, their respective proteins and subcellular localization. It is expected that this will lead to important new insights into disease mechanisms and improved drug discovery strategies to produce novel therapeutics.

Among the major pharmaceutical and biotechnology companies, it is clearly recognized that the business of modern drug discovery is a highly competitive process. All of the many steps involved are inherently complex, and each can involve a high risk of attrition. The players in this business strive continuously to optimize and streamline the process; each seeking to gain an advantage at every step by attempting to make informed decisions at the earliest stage possible. The desired outcome is to accelerate as many key activities in the drug discovery process as possible. This should produce a new generation of robust drugs that offer a high probability of success and reach the clinic and market ahead of the competition.

There has been noticeable emphasis over recent years for companies to aggressively review and refine their strategies to discover new drugs. Central to this has been the introduction and implementation of cutting-edge technologies. Most, if not all, companies have now integrated key technology platforms that incorporate genomics, mRNA expression analysis, relational databases, high-throughput robotics, combinatorial chemistry and powerful bioinformatics. Although it is still early days to quantify the real impact of these platforms in clinical and commercial terms, expectations are high, and it is widely accepted that significant benefits will be forthcoming. This is largely based on data obtained during preclinical studies where the genomic[1,2] and microarray[3,4] technologies have already proved their value.

However, there are several noteworthy outcomes that result from this. Many comments are voiced that scientists armed with these technologies are now commonly faced with data overload. Thus, in some instances, rather than facilitating the decision process, the accumulation of more complex data points, many with unknown consequences, can seem to hinder the process. Also, most drug companies have simultaneously incorporated very similar components of the new technology platforms, the consequence being that it is becoming difficult yet again to determine where a clear competitive advantage will arise. Finally, in recent years, largely as a result of the accessibility of the technologies, there has been an overwhelming emphasis placed on genomic and mRNA data rather than on protein

**Martin J. Page***, **Bob Amess**, **Christian Rohlff**, **Colin Stubberfield** and **Raj Parekh**, Oxford GlycoSciences, 10 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, UK OX14 3YS. *tel: +44 1235 543277, fax: +44 1235 543283, e-mail: martin.page@ogs.co.uk

 1359-6446/99/$ – see front matter  PII: S1359-6446(98)01291-4 

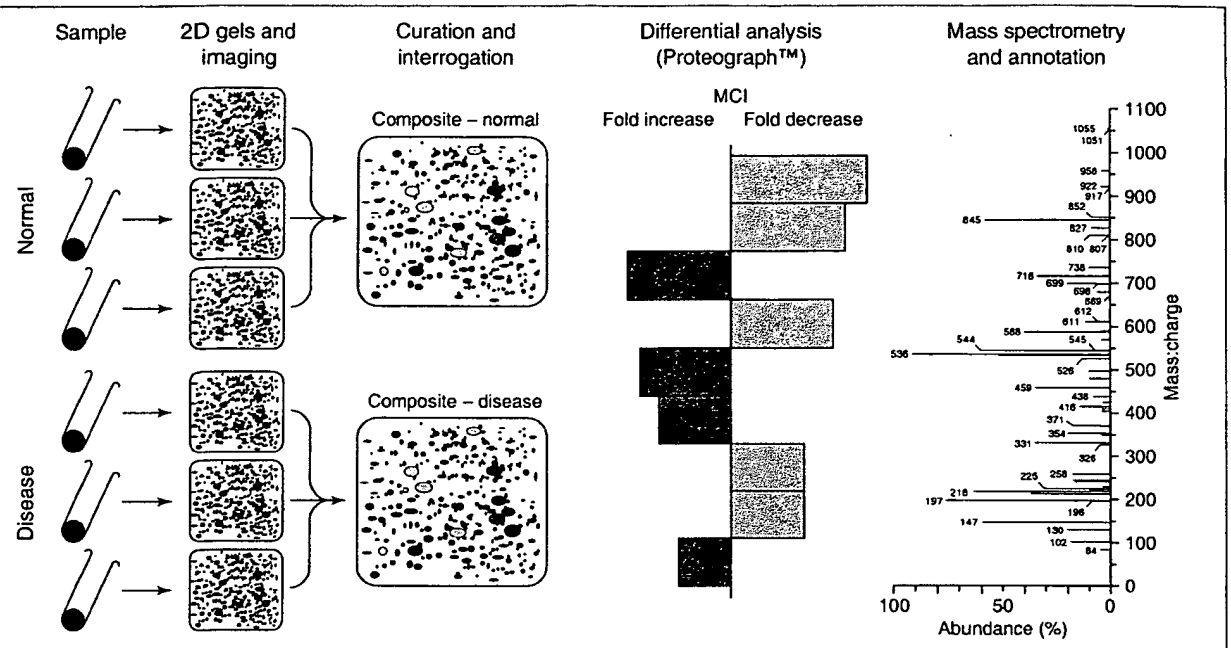


***Figure 1.*** *Steps involved in analysing a biological sample by proteomics. MCI, molecular cluster index.*

analysis. It is important to remember that proteins dictate biological phenotype – whether it is normal or diseased – and are the direct targets for most drugs.

## Proteomics: new technology for the analysis of proteins

It is now timely to recognize that complementary technology in the form of high-throughput analysis of the total protein repertoire of chosen biological samples, namely proteomics, is poised to add a new and important dimension to drug discovery. In a similar fashion to genomics, which aims to profile every gene expressed in a cell, proteomics seeks to profile every protein that is expressed[5–7]. However, there is added information, since proteomics can also be used to identify the post-translational modifications of proteins[8], which can have profound effects on biological function, and their cellular localization. Importantly, proteomics is a technology that integrates the significant advances in two-dimensional (2D) electrophoretic separation of proteins, mass spectrometry and bioinformatics. With these advances it is now possible to consistently derive proteomes that are highly reproducible and suitable for interrogation using advanced bioinformatic tools.

There are many variations whereby different laboratories operate proteomics. For the purpose of this review, the process used at Oxford GlycoSciences (OGS), which uses an industrial-scale operation that is integral to its drug discovery work, will be described. The individual steps of this process, where up to 1000 2D gels can be run and analysed per week, are summarized in Fig. 1. The incoming samples are bar coded and all information relevant to the sample is logged into a Laboratory Information Management System (LIMS) database. There can be a wide range in the type of samples processed, as applicable to individual steps in the drug discovery pipeline, and these will be mentioned later. The samples are separated according to their charge (pI) in the first dimension, using isoelectric focusing, followed by size (MW) using SDS–PAGE in the second dimension. Many modifications have been made to these steps to improve handling, throughput and reproducibility. The separated proteins are then stained with fluorescent dyes which are significantly more sensitive in detection than standard silver methods and have a broader dynamic range. The image of the displayed proteins obtained is referred to as the proteome, and is digitally scanned into databases using proprietary software called ROSETTA™. The images are subsequently curated, which begins with the removal of any artefacts, cropping and the placement of pI/MW landmarks. The images from replicate images are then aligned and matched to one

another to generate a synthetic composite image. This is an important step, as the proteome is a dynamic situation, and it captures the biological variation that occurs, such that even orphan proteins are still incorporated into the analysis.

By means of illustration, Fig. 1 shows the process whereby proteomes are generated from normal and disease samples and how differentially expressed proteins are identified. The potential of this type of analysis is tremendous. For example, from a mammalian cell sample, in excess of 2000 proteins can typically be resolved within the proteome. The quality of this is shown in Fig. 2, which shows representative proteomes from three diverse biological sources: human serum, the pathogenic fungus *Candida albicans* and the human hepatoma cell line Huh7.

### Use of proteomics to identify disease specific proteins

In most cases, the drug discovery process is initiated by the identification of a novel candidate target – almost always a protein – that is believed to be instrumental in the disease process. To date, there is a variety of means whereby drug targets have been forthcoming. These include molecular, cellular and genomic approaches, mostly centred upon DNA and mRNA analysis. The gene in question is isolated, and expression and characterization of its coded protein product – i.e. the drug target – is invariably a secondary event.

With the proteomic approach, the starting point is at the other end of the 'telescope'. Here there is direct and immediate comparison of the proteomes from paired normal and disease materials. Examples of these pairs are: (1) purified epithelial cell populations derived from human breast tumours, matched to purified normal populations of human breast epithelial cells, and (2) the invading pathogenic hyphal form of *C. albicans*, matched to the non-invading yeast form of *C. albicans*. When the proteome images from each pair are aligned, the Proteograph™ software is able to rapidly identify those proteins (each referenced as having a unique molecular cluster index, or MCI) that are either unique, or those that are differentially expressed. Thus, the Proteograph output from this analysis is both qualitative and quantitative.

Proteograph analysis for a particular study can also be undertaken on any number of samples. For example, one might compare anything from a few to several hundred preparations or samples, each from a normal and disease counterpart, and have these analysed in a single Proteograph study. In this way, it is possible to assign strong statistical confidence to the data and in some instances to identify specific subpopulations within the input biological sources. This feature will become increasingly significant in the near future, and there is a clear synergy here whereby proteomics can work closely with pharmacogenomic approaches to stratify patient populations and achieve effective targeted care for the patient. Whatever the source of the materials, the net output of Proteograph analysis is immediate identification of disease specific proteins. This is shown in Fig. 3, which shows the results of a proteograph obtained by comparing untreated human hepatoma cells with cells following exposure to a clinical

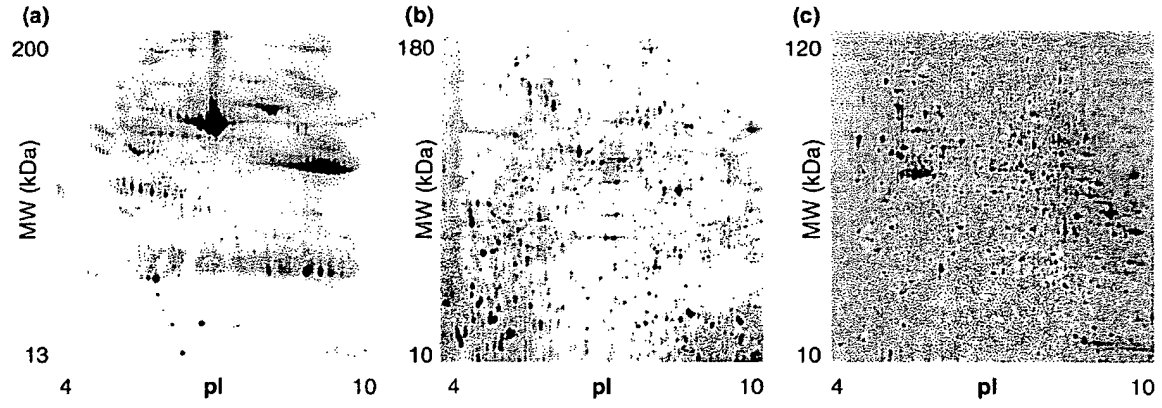


***Figure 2.*** *Representative proteomes obtained from (a) human serum, (b) the pathogenic fungus* Candida albicans *and (c) the human hepatoma cell line Huh7.*

Rosetta

Foregrounds: Huh7 cells treated with 5FU
Backgrounds: Huh7 cells untreated
Upregulated in Huh7 cells treated with 5FU with respect to untreated Huh7 cells
Downregulated in Huh7 cells treated with 5FU with respect to untreated Huh7 cells

| MCI | Fold change | | pI | MW |
|---|---|---|---|---|
| 3589 | 16.3 | | 5.60 | 13718 |
| 4035 | 15.8 | | 8.73 | 13224 |
| 4033 | 12.3 | | 8.94 | 13117 |
| 4052 | 11.7 | | 8.96 | 21434 |
| 4064 | 11.1 | | 9.37 | 29389 |
| 3792 | 10.9 | | 8.75 | 41162 |
| 3976 | 10.5 | | 7.14 | 12061 |
| 3672 | 10.5 | | 8.83 | 24121 |
| 3986 | 10.4 | | 4.61 | 10853 |
| 3759 | 9.9 | | 4.98 | 41420 |
| 4032 | 9.7 | | 4.56 | 13117 |
| 1972 | | −9.4 | 7.92 | 12396 |
| 3587 | 9.4 | | 4.50 | 14954 |
| 2033 | | −9.4 | 7.93 | 11126 |
| 3984 | 8.9 | | 5.32 | 11090 |
| 2403 | | −8.8 | 6.04 | 25950 |
| 3748 | 8.8 | | 6.32 | 35513 |
| 2105 | | −8.7 | 4.76 | 20803 |
| 3897 | 8.7 | | 4.94 | 87842 |
| 4221 | 8.6 | | 5.78 | 71963 |

***Figure 3.*** *Table of differential protein expression profiles, referred to as a Rosetta Proteograph™, between Huh7 cells with and without the cytotoxic agent 5-FU. Bars are quantized and do not represent exact fold change values.*

cytotoxic agent. In this instance, only the top 20 differentially expressed MCIs are shown, but the readout would normally extend to a defined cut-off value, typically a two-fold or greater difference in expression levels, determined by the user.

In a typical analysis involving disease and normal mammalian material, in which each proteome would have ~2000 protein features each assigned an MCI, the proteograph might identify somewhere in the region of 50–300 MCIs that are unique or differentially expressed. To capitalize rapidly on these data, at OGS a high-throughput mass spectrometry facility coupled to advanced databases to annotate these MCIs as individual proteins is applied. As these are all disease specific proteins, each could represent a novel target and/or a novel disease marker. The process becomes even more powerful when a panel of features, rather than individual features, are assigned. The relevance of this is apparent when one considers that most diseases, if not all, are multifactorial in nature and arise from polygenic changes. Rather than analysing events in isolation, the ability to examine hundreds or thousands of events simultaneously, as shown by proteomics, can offer real advantages.

### *Identification and assignment of candidate targets*

The rapid identification and assignment of candidate targets and markers represents a huge challenge, but this has been greatly facilitated by combining the recent advances made in proteomics and analytical mass spectrometry[9]. Using automated procedures it is now possible to annotate proteins present in femtomole quantities, which would depict the low abundance class of proteins. The process of annotation is similarly aided by the quality and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals). In this respect, the advances in proteomics have benefited considerably from the breakthroughs achieved with genomics.

From an application perspective, cancer studies provide a good opportunity whereby proteomics can be instrumental in identifying disease specific proteins, because it is often feasible to obtain normal and diseased tissue from the same patient. For example, proteomic studies have been reported on neuroblastomas[10], human breast proteins from normal and tumour sources[11–13], lung tumours[14], colon tumours[15] and bladder tumours[16]. There are also proteomic studies reported within the cardiovascular therapeutic area, in which disease or response proteins are identified[17,18].

Genomic microarray analysis can similarly identify unique species or clusters of mRNAs that are disease specific. However, in some instances, there is a clear lack of correlation between the levels of a specific mRNA and its corresponding protein (Ref. 19, Gypi, S.P. *et al.*, submitted). This has now been noted by many investigators and reaffirms that post-transcriptional events, including protein stability, protein modification (such as phosphorylation, glycosylation, acylation and methylation) and cell localization, can constitute major regulatory steps. Proteomic analysis captures all of these steps and can therefore provide unique and valuable information independent from, or complementary to, genomic data.

### Proteomics for target validation and signal transduction studies

The identification of disease specific proteins alone is insufficient to begin a drug screening process. It is critical to assign function and validation to these proteins by confirming they are indeed pivotal in the disease process. These studies need to encompass both gain- and loss-of-function analyses. This would determine whether the activity of a candidate target (an enzyme, for example), eliminated by molecular/cellular techniques, could reverse a disease phenotype. If this happened, then the investigator would have increased confidence that a small-molecule inhibitor against the target would also have a similar effect. The proposal of candidate drug targets is often not a difficult process, but validating them is another matter. Validation represents a major bottleneck where the wrong decision can have serious consequences[20].

Proteomics can be used to evaluate the role of a chosen target protein in signal transduction cascades directly relevant to the disease. In this manner, valuable information is forthcoming on the signalling pathways that are perturbed by a target protein and how they might be corrected by appropriate therapeutics. Techniques that are well established in one-dimensional protein studies to investigate signalling pathways, such as western blotting and immunoprecipitation, are highly suited to proteomic applications. For example, the proteomes obtained can be blotted onto membranes and probed with antibodies against the target protein or related signalling molecules[21–23]. Because proteomics can resolve >2000 proteins on a single gel, it is possible to derive important information on specific isoforms (such as glycosylated or phosphorylated variants) of signalling molecules. This will result in characterization of how they are altered in the disease process. Western immunoblotting techniques using high-affinity antibodies will typically identify proteins present at ~10 copies per cell (~1.7 fmol); this is in contrast to the best fluorescent dyes currently available that are limited to imaging proteins at 1000 or more copies per cell. The level of sensitivity derived by these applications will greatly facilitate interpretation of complex signalling pathways and contribute significantly to validation of the target under study.

#### *Immunoprecipitation studies*

Similarly, immunoprecipitation studies are another useful way to exploit the resolving power of proteomics[24,25]. In this instance, very large quantities of protein (e.g. several milligrams) can be subjected to incubation with antibodies against chosen signalling molecules. This allows high-affinity capture of these proteins, which can subsequently be eluted and electrophoresed on a 2D gel to provide a high-resolution proteome of a specific subset of proteins. Detection by blot analysis allows the identification of extremely small amounts of defined signalling molecules. Again, the different isoforms of even very low abundance proteins can be seen, and, very importantly, the technique allows the investigator to identify multiprotein complexes or other proteins that co-precipitate with the target protein. These coassociating proteins frequently represent signalling partners for the target protein, and their identification by mass spectrometry can lead to invaluable information on the signalling processes involved.

The depth of signal transduction analysis offered by proteomics, and the utility for target validation studies, can be extended even further by applying cell fractionation studies[26–28]. By purifying subcellular fractions, such as membrane, nuclear, organelle and cytosolic, it is possible to assign a localization to proteins of interest and to follow their trafficking in a cell. Enrichment of these fractions will also allow much higher representation of low abundance proteins on the proteome. Their detection by fluorescent dyes or immunoblot techniques will lead to the identification of proteins in the range of 1–10 copies per cell, putting the sensitivity on a par with genomic approaches.

These signal transduction analyses can be of additional value in experiments where inhibitors derived from a screening programme against the target are being evaluated for their potency and selectivity. The inhibitors can encompass small molecules, antisense nucleic acid constructs, dominant-negative proteins, or neutralizing antibodies microinjected into cells. In each case, proteome analysis can provide unique data in support of validation studies for a chosen candidate drug target.

### Proteomics and drug mode-of-action studies

Once a validated target is committed to a screening regimen to identify and advance a lead molecule, it is important to confirm that the efficacy of the inhibitor is through the expected mechanism. Such mode-of-action studies are usually tackled by various cell biological and biochemical methods. Proteomics can also be usefully applied to these studies and this is illustrated below by describing data obtained with OGT719. This is a novel galactosyl derivative of the cytotoxic agent 5-fluorouracil (5-FU), which is currently being developed by OGS for the treatment of hepatocellular carcinoma and colorectal metastases localized in the liver. The premise underpinning the design and rationale of OGT719 was to derive a 5-FU prodrug capable

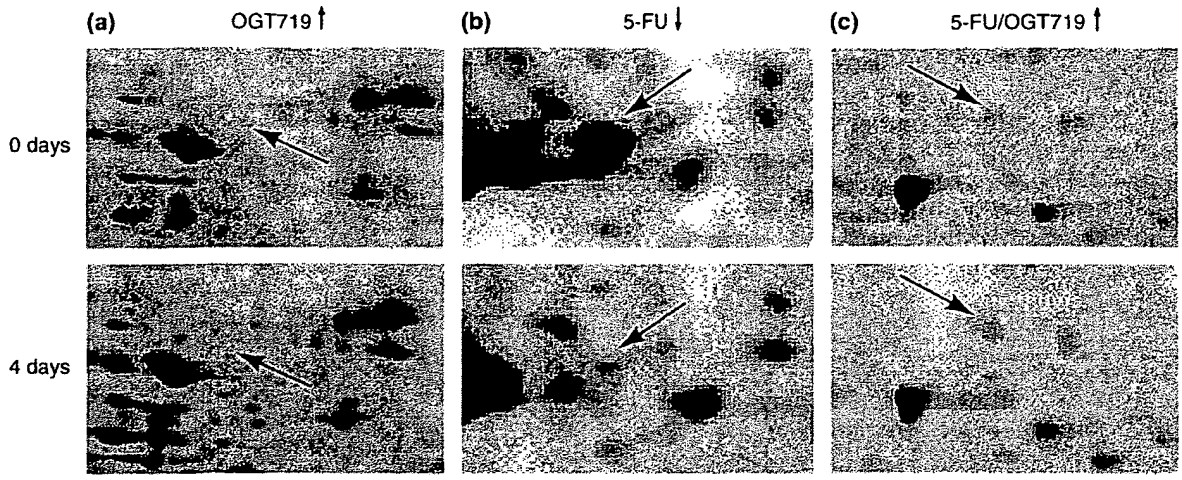


***Figure 4.*** *Features that are specifically up- or downregulated in Huh7 cells by either 5-fluorouracil (5-FU) or OGT719: (a) elongation factor 1α2, (b) novel (three peptides by MS-MS) and (c) α-subunit of prolyl-4-hydroxylase. Arrows indicate up- or downregulated.*

of targeting, and being retained in, cells bearing the asialoglycoprotein receptor (ASGP-r), including hepatocytes[29], hepatoma Huh7 cells[30] and some colorectal tumour cells[31]. The growth of the human hepatoma cell line Huh7 is inhibited by 5-FU or by OGT719. If the inhibition by OGT719 were the result of uptake and conversion to 5-FU as the active component, then it would be expected that Huh7 cells would show similar proteome profiles following exposure to either drug.

To examine these possibilities, we conducted an experiment taking samples of Huh7 cells that had been treated with $IC_{50}$ doses of either OGT719 or 5-FU. Total cell lysates were prepared and taken through 2D electrophoresis, fluorescence staining, digital imaging and Proteograph analysis. To facilitate the interpretation of the data across all of the 2291 features seen on the proteomes, drug-induced protein changes of fivefold or greater, identified by the Proteograph, were analysed further. Interestingly, from this analysis 19 identical proteins were changed fivefold or more by both drugs, strongly suggesting similarities in the mode of action for these two compounds.

Thus, from very complex data involving >2000 protein features, using proteomics it is possible to analyse quantitatively and qualitatively each protein during its exposure to drugs. The biologist is now able to focus a series of further studies specifically on an enriched subset of proteins. Figure 4 shows highlighted examples of the selected areas of the proteome where some of these identified proteins in the above study are altered in response to either or both drugs.

Several of the proteins identified above as being modulated similarly by 5-FU or OGT719 in Huh7 cells were subjected to tandem mass-spectrometric analysis for annotation. Some of these, such as the nuclear ribosomal RNA-binding protein[32], can be placed into pyrimidine pathways or related cell cycle/growth biochemical pathways in which 5-FU is known to act.

To attribute further significance to the proteome mode-of-action studies with OGT719, another cell line, the rat sarcoma HSN, was used. Growth of these cells is inhibited by 5-FU, but they are completely refractory to OGT719; notably they lack the ASGP-r, which might explain this finding (unpublished). For our proteome studies, HSN cells were treated with 5-FU or OGT719 over a time course of one, two and four days. At each time point, cells were harvested and processed to derive proteomes and Proteographs. As before, we purposely focused on those proteins that increased or decreased by fivefold or more. In this instance, there were no proteins co-modulated by the two drugs. This is perhaps to be expected, given that the HSN cells are killed by 5-FU and yet are refractory to OGT719.

*Clear potential*

The above is just an example of how proteomics can be used to address the mode of action of anticancer drugs. The potential of this approach is clear, and one can envisage situations where it will be profitable to compare the proteomes of cells in which the drug target has been eliminated by molecular knockout techniques, or with small-molecule inhibitors believed to act specifically on the same target. In addition to using proteomics to examine the action of drugs, it is also possible to use this approach to gauge the extent of nonspecific effects that might eventually lead to toxicity. For instance, in the example used above with HSN cells treated with OGT719, although cell growth was not affected, the levels of several specific proteins were changed. Further investigation of these proteins and the signalling pathways in which they are involved could be illuminating in predicting the likelihood or otherwise of long-term toxicity.

## Use of proteomics in formal drug toxicology studies

A drug discovery programme at the stage where leads have been identified and mode-of-action studies are advanced, will proceed to investigate the pharmacokinetic and toxicology profile of those agents. These two parameters are of major importance in the drug discovery process, and many agents that have looked highly promising from *in vitro* studies have subsequently failed because of insurmountable pharmacokinetic and/or toxicity problems *in vivo*. Whereas the pharmacokinetic properties of a molecule can now be characterized quickly and accurately, toxicity studies are typically much longer and more demanding in their interpretation.

The ability to achieve fast and accurate predictions of toxicity within an *in vivo* setting would represent a big step forward in accelerating any drug discovery programme. Toxicity from a drug can be manifested in any organ. However, because the liver and kidney are the major sites in the body responsible for metabolism and elimination of most drugs, it is informative to examine these particular organs in detail to provide early indications about events that might result in toxicity.

The basis for most xenobiotic metabolizing activity is to increase the hydrophilicity of the compound and so facilitate its removal from the body. Most drugs are metabolized in the liver via the cytochrome P450 family of enzymes, which are known to comprise a total of ~200 different members[33,34], encompassing a wide array of overlapping specificities for different substrates. In addition to clearance, they also play a major role in metabolism that can lead to the production and removal of toxic species, and in some instances it is possible to correlate the ability or failure to remove such a toxin with a specific P450 or subgroup.

*Unique P450 profiles*

Each individual person will have a slightly different P450 profile, largely from polymorphisms and changes in expression levels, although other genetic and environmental factors aside from P450 also need to be taken into consideration. A significant amount of research is currently being directed towards this field – known as pharmacogenomics – with the aim of predicting how a patient will respond to a drug, as determined by their genetic make-up[35–37]. The marked variation of individuals in their ability to clear a compound can be one of the key factors in deciding the overall pharmacokinetic profile of a drug. Not only will this have a bearing on the likelihood of a patient responding to a treatment, but it will also be a factor in determining the possibility of their experiencing an adverse effect.

Many pharmaceutical companies are already employing genomic approaches, involving P450 measurements, as a key step in their assessment of the toxicological profile of a candidate drug and therefore of its suitability, or otherwise, to be considered for human clinical trials. There are limits to this approach, however. Whereas the P450 mRNA profiling can predict with some accuracy the likely metabolic fate of a drug, it will not provide information on whether the metabolites would subsequently lead to toxicity. Besides the patient-to-patient differences in steady-state levels of the P450s, there are also characteristic induction responses of these enzymes to some drugs. Moreover, as there can be some doubt over the correlation of mRNA levels and the corresponding protein levels, there is scope for misinterpretation of the results and hence real advantages to be gained from a proteome approach. In both instances, the ability to examine entire proteome profiles, including the P450 proteins, will be a significant advantage in understanding and predicting the metabolism and toxicological outcome of drugs.

In addition to direct organ and tissue studies, the serum, which collects the majority of toxicity markers released from susceptible organs and tissues throughout the entire body, can be utilized. Serum is rich in nuclease activity and, as pharmacogenomics is not suited to deal with these samples, valuable markers of toxicity could go undetected. However, by using proteomics for these types of analyses, serum markers (and clusters thereof) are now accessible for evaluation as indicators of toxicity.

*Pharmacoproteomics*

Proteomics can thus be used to add a new sphere of analysis to the study of toxicity at the protein level, and in the era of '-omics' there is a case to be made to adopt the term 'Pharmacoproteomics™'. Animals can be dosed with increasing levels of an experimental drug over time, and serum samples can be drawn for consecutive proteome analyses. Using this procedure, it should be possible to identify individual markers, or clusters thereof, that are dose related and correlate with the emergence and severity of toxicity. Markers might appear in the serum at a defined drug dose and time that are predictive of early toxicity within certain organs and if allowed to continue will have damaging consequences. These serum markers could subsequently be used to predict the response of each individual and allow tailoring of therapy whereby optimal efficacy is achieved without adverse side effects being apparent. This application can obviously extend to tracking toxicity of drugs in clinical trials where serum can be readily drawn and analysed. Surrogate markers for drug efficacy could also be detected by this procedure and could facilitate the challenge of identifying patient classes who will respond favourably to a drug and at what dosage.

## Conclusions

By contrast to the agents administered to patients in clinical wards, the process of drug discovery is not a prescriptive series of steps. The risks are high and there are long timelines to be endured before it is known whether a candidate drug will succeed or fail. At each step of the drug discovery process there is often scope for flexibility in interpretation, which over many steps is cumulative. The pharmaceutical companies most likely to succeed in this environment are those that are able to make informed accurate decisions within an accelerated process.

The genomics revolution has impacted very positively upon these issues and now has a powerful new partner in proteomics. The ability to undertake global analysis of proteins from a very wide diversity of biological systems and to interrogate these in a high-throughput, systematic manner will add a significant new dimension to drug discovery. Each step of the process from target discovery to clinical trials is accessible to proteomics, often providing unique sets of data. Using the combination of genomics and proteomics, scientists can now see every dimension of their biological focus, from genes, mRNA, proteins and their subcellular localization. This will greatly assist our understanding of the fundamental mechanistic basis of human disease and allow new improved and speedier drug discovery strategies to be implemented.

## REFERENCES


1 Crooke, S.T. (1998) *Nat. Biotechnol.* 16, 29–30
2 Dykes, C.W. (1996) *Br. J. Clin. Pharmacol.* 42, 683–695
3 Schena, M. *et al.* (1998) *Trends Biotechnol.* 16, 301–306
4 Ramsay, G. (1998) *Nat. Biotechnol.* 16, 40–44
5 Anderson, N.L. and Anderson, N.G. (1998) *Electrophoresis* 19, 1853–1861
6 James, P. (1997) *Biochem. Biophys. Res. Commun.* 231, 1–6
7 Wilkins, M.R. *et al.* (1996) *Biotechnol. Genet. Eng. Rev.* 13, 19–50
8 Parekh, R.B. and Rohlff, C. (1997) *Curr. Opin. Biotechnol.* 8, 718–723
9 Figeys, D. *et al.* (1998) *Electrophoresis* 19, 1811–1818
10 Wimmer, K. *et al.* (1996) *Electrophoresis* 17, 1741–1751
11 Giometti, C.S., Williams, K. and Tollaksen, S.L. (1997) *Electrophoresis* 18, 573–581
12 Williams, K. *et al.* (1998) *Electrophoresis* 19, 333–343
13 Rasmussen, R.K. *et al.* (1998) *Electrophoresis* 19, 818–825
14 Hirano, T. *et al.* (1995) *Br. J. Cancer* 72, 840–848
15 Ji, H. *et al.* (1997) *Electrophoresis* 18, 605–613
16 Ostergaard, M. *et al.* (1997) *Cancer Res.* 57, 4111–4117
17 Patel, V.B. *et al.* (1997) *Electrophoresis* 18, 2788–2794
18 Arnott, D. *et al.* (1998) *Anal. Biochem.* 258, 1–18
19 Anderson, L. and Seilhamer, J. (1997) *Electrophoresis* 18, 533–537
20 Rastan, S. and Beeley, L.J. (1997) *Curr. Opin. Genet. Dev.* 7, 777–783
21 Gravel, P. *et al.* (1995) *Electrophoresis* 16, 1152–1159
22 Qian, Y. *et al.* (1997) *Clin. Chem.* 43, 352–359
23 Sanchez, J.C. *et al.* (1997) *Electrophoresis* 18, 638–641
24 Watts, A.D. *et al.* (1997) *Electrophoresis* 18, 1086–1091
25 Asker, N. *et al.* (1995) *Biochem. J.* 308, 873–880
26 Ramsby, M.L., Makowski, G.S. and Khairallah, E.A. (1994) *Electrophoresis* 15, 265–277
27 Huber, L.A. (1995) *FEBS Lett.* 369, 122–125
28 Corthals, G.L. *et al.* (1997) *Electrophoresis* 18, 317–323
29 Hubbard, A.L., Wall, D.A. and Ma, A. (1983) *J. Cell Biol.* 96, 217–229
30 Zeng, F.Y., Oka, J.A. and Weigel, P.H. (1996) *Biochem. Biophys. Res. Commun.* 218, 325–330
31 Mu, J-Z. *et al.* (1994) *Biochim. Biophys. Acta* 1222, 483–491
32 Ghoshal, K. and Jacob, S.T. (1997) *Biochem. Pharmacol.* 53, 1569–1575
33 Guengerich, F.P. and Parikh, A. (1997) *Curr. Opin. Biotechnol.* 8, 623–628
34 Rendic, S. and Di Carlo, F.J. (1997) *Drug Metab. Rev.* 29, 413–580
35 Vermes, A., Guchelaar, H.J. and Koopmans, R.P. (1997) *Cancer Treat. Rev.* 23, 321–339
36 Housman, D. and Ledley, F.D. (1998) *Nat. Biotechnol.* 16, 492–493
37 Persidis, A. (1998) *Nat. Biotechnol.* 16, 209–210

Letter

# Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins

Hedi Hegyi and Mark Gerstein[1]

*Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA*

Annotation transfer is a principal process in genome annotation. It involves "transferring" structural and functional annotation to uncharacterized open reading frames (ORFs) in a newly completed genome from experimentally characterized proteins similar in sequence. To prevent errors in genome annotation, it is important that this process be robust and statistically well-characterized, especially with regard to how it depends on the degree of sequence similarity. Previously, we and others have analyzed annotation transfer in single-domain proteins. Multi-domain proteins, which make up the bulk of the ORFs in eukaryotic genomes, present more complex issues in functional conservation. Here we present a large-scale survey of annotation transfer in these proteins, using scop superfamilies to define domain folds and a thesaurus based on SWISS-PROT keywords to define functional categories. Our survey reveals that multi-domain proteins have significantly less functional conservation than single-domain ones, except when they share the exact same combination of domain folds. In particular, we find that for multi-domain proteins, approximate function can be accurately transferred with only 35% certainty for pairs of proteins sharing one structural superfamily. In contrast, this value is 67% for pairs of single-domain proteins sharing the same structural superfamily. On the other hand, if two multi-domain proteins contain the same combination of two structural superfamilies the probability of their sharing the same function increases to 80% in the case of complete coverage along the full length of both proteins, this value increases further to > 90%. Moreover, we found that only 70 of the current total of 455 structural superfamilies are found in both single and multi-domain proteins and only 14 of these were associated with the same function in both categories of proteins. We also investigated the degree to which function could be transferred between pairs of multi-domain proteins with respect to the degree of sequence similarity between them, finding that functional divergence at a given amount of sequence similarity is always about two-fold greater for pairs of multi-domain proteins (sharing similarity over a single domain) in comparison to pairs of single-domain ones, though the overall shape of the relationship is quite similar. Further information is available at http://partslist.org/func or http://bioinfo.mbb.yale.edu/partslist/func.

The ultimate goal of the genome projects is to determine the structure and function of all the newly identified gene products. Fundamentally, this will be carried out via annotation transfer, transferring the structural and functional annotation from an experimentally characterized protein (as in a model organism such as *Escherichia coli*) to a predicted protein in a newly sequenced genome that shares similarity in sequence. The degree of annotation transferred will depend on the degree of sequence similarity. This process is shown schematically in Figure 1. In this paper, we aim to address this major question in bioinformatics, specifically focusing on multi-domain proteins, as they make up the bulk of the proteome in eukaryotic organisms (Gerstein 1998).

Our work is a direct outgrowth of two previous analyses of ours that concentrated on single-domain proteins. In an earlier paper, we found that the different structural classes of the scop classification system have different propensities to carry out certain types of function (Hegyi and Gerstein 1999). In particular, while the alpha/beta folds were disproportionately associated with enzymes and all-alpha and small folds with non-enzymes, the alpha + beta structures had an equal tendency for both enzymatic and non-enzymatic functions.

[1]Corresponding author.
E-MAIL Mark.Gerstein@yale.edu
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.183801.

Wilson et al. (2000) compared a large number of protein domains to one another in a pair-wise fashion with respect to similarities in sequence, structure, and function. Using a hybrid functional classification scheme merging the ENZYME and FlyBase systems (Gelbart et al. 1997; Bairoch 2000), they found that precise function is not conserved below 30–40% identity, although the broad functional class is usually preserved for sequence identities as low as 20–25%, given that the sequences have the same fold. Their survey also reinforced the previously established general exponential relationship between structural and sequence similarity (Chothia and Lesk 1986).

## Other Work on Establishing Relationships between Sequence, Structure, and Function

Several other groups have studied the relationship between sequence, structure, and function in detail, attempting to determine the extent to which functional transference between matching proteins is feasible (Shah and Hunger 1997; Martin et al. 1998; Thornton et al. 1999, 2000; Zhang et al. 1999; Shapiro and Harris 2000; Todd et al. 2001). Orengo et al. (1999) analyzed protein families in the CATH database and concluded that > 96% of the folds in the PDB are associated with a single homologous family. By investigating enzymatic folds they also found that more than 95% of homologous families show either single or closely related functions.


11:1632–1640 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

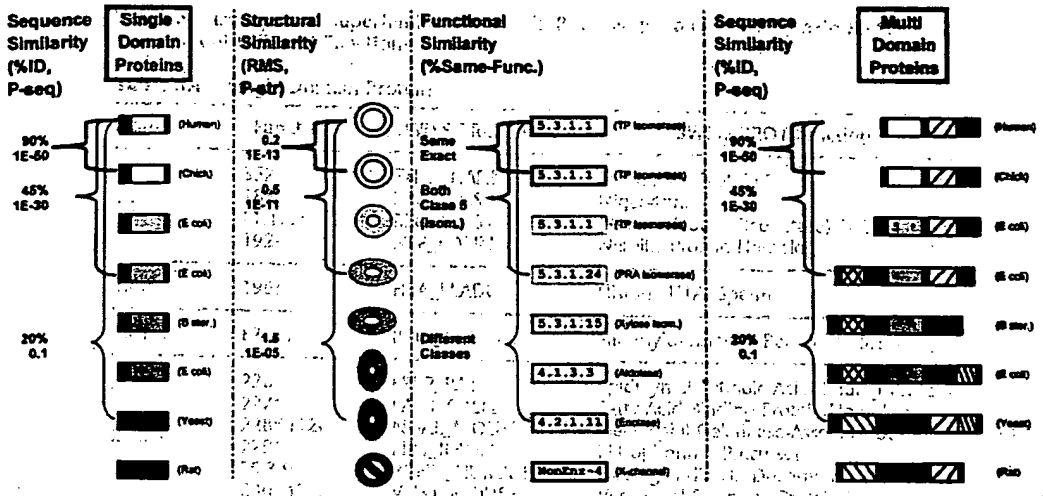


**Figure 1** Schematic illustrating annotation transfer. This figure illustrates the process of annotation transfer for a group of hypothetical TIM barrel proteins. The *leftmost* panel represents sequence comparisons between idealized barrel domains from a number of organisms. The next panel shows analogous results for structural comparison, and the panel after that, functional comparison. The *rightmost* panel represents sequence comparisons between idealized multi-domain proteins that match over a single domain, the subject of much of this paper.

Pawlowski et al. (2000) studied the relationship between sequence and functional similarity in the twilight zone of 10%–15% sequence similarity and found a clear correlation between the two, with functional similarity based on the E.C. classification of enzymes.

Russell et al. (1997) analyzed binding sites in proteins with similar 3D structures and estimated that 90% of new remote homolog have common binding sites and similar functions. Eisenstein et al. (2000) evaluated the first results from the structural genomics projects and found that in many instances the protein structure itself offers an important clue to its biological function. Stawiski et al. (2000) found that function could be predicted rather successfully for just the proteases. Devos and Valencia (2000) presented a critical view of function transference between similar sequences, highlighting the limitations of this process due to errors in databases and the inherent complexity of the relationship between protein sequence-structure and function that does not allow "simplistic interpretations." They also found that binding sites are the least conserved features between related proteins while the catalytic activity of enzymes is the most conserved one.

## Multi-Domain Proteins with Divergent Functions: How Common?

Most of these previous investigations focused on single-domain proteins or did not distinguish between single- and multi-domain ones. It is not clear how the multi-domain proteins with various functions behave with respect to functional conservation; namely, whether they are more or less conserved than their single-domain counterparts. In particular, as shown in Figure 1, if one multi-domain protein shares a single domain fold with another one, it is not clear the degree to which the functional conservation of these proteins is constrained by the shared part, and to what degree it is influenced by other domains that are not shared.

Specific groups of proteins that have the same combination of structural domains but dramatically different functions illustrate this situation. One example is the combination of the SH3-domain (scop superfamily identifier 2.24.2) and the P-loop containing NTP hydrolase (3.29.1). While in higher organisms this combination is associated with presynaptic and tumor suppressor functions (SWISS-PROT names SP02_HUMAN and DLGI_DROME, respectively), in the lower Dictyostelium it was found in myosin (MYSP_DICDI). Another example is the combination of the FAD/NAD(P)-binding superfamily and FAD-linked reductases C-terminal superfamily (3.4.1 and 4.12.1 superfamilies, respectively). In one group of proteins they appear in enzymes of the oxidoreductase group (e.g. OXDA_CAEEL or PHHY_PSEAE), while in another they are found in a dissociation inhibitor (e.g. GDIA_HUMAN). It should be noted that the proteins are not covered completely by the structural matches, so it is quite possible that the rest of them contain totally different domains that are responsible for the dramatically different functions. However, do these two examples show a rather rare or a more frequent phenomenon? How often do multi-domain proteins, sharing the same structural domain composition, differ in their functions?

In this paper, we attempt to provide a comprehensive answer to this question. This is particularly timely given that most of the unknown proteins in eukaryotic genomes are multi-domain. We use the same approach as in our previous analyses, comparing the sequences of the structural domains in scop to those of SWISS-PROT using BLASTP. We focus on the functional divergence of single and multi-domain proteins, extending previous investigations of single-domain proteins. Also, in comparison to previous work, we focus more on non-enzymatic functions and scop structural superfamilies, instead of folds.

# RESULTS

## Our Approach to Functional and Structural Assignment

We used the BLASTP program (version 2.0) (Altschul et al. 1997) to identify the scop 1.39 (Murzin et al. 1995) structural domains in SWISS-PROT (version 37) (Bairoch and Apweiler

2000) with $e = 10^{-4}$. We removed the hypothetical and fragment proteins. This resulted in two sets of proteins.

### *Single-Domain*

Of the single-domain matches, only those that were almost completely covered with a match to a single structural domain were selected. (The maximum number of uncovered residues was set at 70 with an additional condition that a maximum of 40 residues on the N-terminal end and 30 residues on the C-terminus were allowed to be uncovered.) These criteria resulted in 1818 single-domain proteins being selected from SWISS-PROT.

### *Multi-Domain*

We selected 4763 multi-domain proteins from SWISS-PROT. All of these matched (in different locations) at least two domains of known structure belonging to different scop superfamilies (see schematic in Figure 1). We also selected a subset of these proteins that have almost their entire length covered by matches with structural domains (allowing again a maximum of 70 uncovered residues). This selection resulted in 2829 proteins being selected from SWISS-PROT. (In all cases, duplicate matches were removed, i.e., a protein at a certain location matches only one structural domain.)

We set out to compare these two sets of proteins for functional divergence. As previously, we divided functions into enzyme and non-enzyme (Hegyi and Gerstein 1999). Enzymatic functions were classified by the EC system (Bairoch 2000). Comparisons of enzymatic functions were treated the same way as in our earlier analyses, that is, if they differ in the first three components of their respective EC numbers, they were considered different. This implied that our analysis dealt with a total of 112 enzymatic functions. Non-enzymatic functions were classified into 508 different categories based on a simple thesaurus we assembled of synonymous keywords drawn from SWISS-PROT description lines. In addition, we created 49 categories for functions that have an enzymatic component but which are not part of the EC system. This gave us a total of 669 functions (112 + 508 + 49). (The list of all the functional categories is described further in Table 2 below, and also can be found on the Web at http://bioinfo.mbb.yale.edu/partslist/func or http://partslist.org/func.)

## Overall Distribution of the Matches

Figure 2 shows the most commonly observed multi-domain combinations in a set of recently sequenced genomes. The occurrences of further combinations are available from the Web site. Clearly, the distribution is very skewed, with certain combinations, such as 3.29–2.32, and 2.29–4.61 tending to predominate.

Figure 3 shows the overall distribution of the single-domain and multi-domain matches in the different structural classes. The distribution of matches between enzymes and non-enzymes in multi-domain proteins largely agrees with that in the single-domain proteins. The multi-domain matches follow the overall tendency of the alpha/beta folds to be associated with enzymes to a larger extent and the all-alpha and small folds with non-enzymes. However, the values for the multi-domain matches are generally less extreme than for single-domains; for example, the 10-fold difference between single-domain alpha/beta enzymes and non-enzymes decreases to about twofold in multi-domain proteins. Another significant difference is the reduction in the number of multi-domain non-enzymes in the all-beta and alpha + beta structural classes compared to the single-domain matches. Altogether, there are more enzymes than non-enzymes among the multi-domain proteins (2805 enzymes vs. 1958 non-enzymes) whereas for single-domain proteins, the opposite is true (850 enzymes vs. 968 non-enzymes).

| fold 1 | fold 2 | aful | aeol | mthe | phor | aeae | cele | aeae | dme | ecol | bsub | mtub | hinf | hpyl | mgen | mpne | mbur | tpal | ctra | cpne | cpne | rpro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.29 | 2.32 | 4 | 3 | 4 | 3 | 12 | 14 | 6 | 7 | 8 | 4 | 6 | 7 | 6 | 3 | 3 | 4 | 6 | 3 | 4 | 4 | |
| 2.29 | 8.61 | 1 | 1 | -1 | 2 | 6 | 3 | 2 | 4 | 6 | 4 | 4 | 3 | 3 | 3 | 4 | 1 | 2 | 3 | 3 | 2 | |
| 4.1 | 4.34 | 1 | 1 | 1 | 1 | 5 | 3 | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | |
| 1.28 | 3.29 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 3.4 | 4.48 | 4 | 1 | 1 | 2 | 3 | 4 | 1 | 2 | 4 | 3 | 5 | 2 | | 2 | 2 | 1 | 1 | 1 | 1 | 2 | |
| 3.22 | 4.42 | 1 | 1 | 1 | 0 | 4 | 5 | 3 | 4 | 5 | 4 | 4 | 3 | 4 | 1 | 1 | 2 | 2 | 3 | 3 | 1 | |
| 2.32 | 4.1 | 1 | 1 | 1 | 1 | 4 | 2 | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | | 1 | 1 | |
| 2.32 | 2.33 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 4.32 | 3.1 | 1 | 1 | 1 | 2 | 5 | 1 | 1 | 1 | 4 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | |
| 3.23 | 4.89 | 3 | 3 | 3 | 0 | 9 | 10 | 6 | 6 | 6 | 8 | 7 | 2 | 4 | 0 | 0 | 1 | 1 | 2 | 2 | 2 | |
| 3.47 | 5.17 | 0 | 0 | 1 | 0 | 12 | 10 | 1 | 3 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | |
| 4.72 | 5.13 | 1 | 0 | 0 | 0 | 1 | 3 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | |
| 3.22 | 8.1 | 1 | 1 | 1 | 0 | 8 | 3 | 2 | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | |
| 3.5 | 3.1 | 1 | 1 | 2 | 1 | | | 1 | 1 | 5 | 1 | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 4.61 | 3.62 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 2 | 2 | 1 | 1 | 1 | 0 | 0 |
| 1.76 | 3.3 | 1 | 0 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | |
| 4.29 | 4.1 | 1 | | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | | 1 | | 0 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 2.32 | 4.34 | | | | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 3.22 | 1.79 | 1 | 1 | 1 | 0 | 3 | 1 | 2 | 2 | 2 | 4 | 3 | 2 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | |
| 3.52 | 2.36 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |

**Figure 2** Distribution of multi-domain combinations amongst the genomes. The figure shows the occurrence of multi-domain fold combinations in a number of genomes, indicating its great variability. Each row indicates a particular combination of scop fold pairs (using scop 1.39), where a fold pair is defined as two distinct folds occurring in tandem in a protein. Each column represents a different genome, using the four-letter codes in the PartsList system (Qian et al. 2001): Aaeo, *Aquifex aeolicus*; Aful, *Archaeoglobus fulgidus*; Bbur, *Borrelia burgdorferi*; Bsub, *Bacillus subtilis*; Cele, *Caenorhabditis elegans*; Cpne, *Chlamydia pneumoniae*; Ctra, *Chlamydia trachomatis*; Ecol, *Echerischia coli*; Hinf, *Haemophilus influenzae* Rd; Hpyl, *Helicobacter pylori*; Mthe, *Methanobacterium thermoautotrophicum*; Mjan, *Methanococcus jannaschii*; Mtub, *Mycobacterium tuberculosis*; Mgen, *Mycoplasma genitalium*; Mpne, *Mycoplasma pneumoniae*; Phor, *Pyrococcus horikoshii*; Rpro, *Rickettsia prowazekii*; Scer, *Saccharomyces cerevisiae*; Syne, *Synechocystis* sp.; Tpal, *Treponema pallidum*. The numbers in each intersection cell indicate the number of times the fold pairs occur in a genome. Only the 20 most common fold pair combinations are shown here; the remainder are shown on the Web site (http://partslist.org/func). If a cell is greater than 6, it is shaded black; between 3 and 6, gray; and below 3, white. The blank spaces show instances in which one of the pairs does not occur in the organism at all (indicated by a value of -1 in the data table on the Web site). The fold assignments are done in a fashion consistent with those in PartsList and associated systems (Gerstein 1997; Lin et al. 2000; Drawid et al. 2001; Harrison et al. 2001; Qian et al. 2001).

Table 1 summarizes the distribution of superfamilies and superfamily combinations among the major functional classes, i.e. whether they have only enzymatic, only non-enzymatic or both enzymatic and non-enzymatic functionality. Altogether, 215 superfamilies were found in single-domain proteins and 310 in multi-domain ones. As 70 superfamilies were found in both, altogether 455 distinct structural superfamilies matched a SWISS-PROT protein with our required coverage criteria (described above). Similarly, we apportioned the 281 superfamily combinations observed in multi-domain proteins amongst different broad functional categories.

In single-domain proteins there are about as many superfamilies with exclusively enzymatic functionality as there are those with exclusively non-enzymatic functions (82 vs. 78). In contrast, in multi-domain proteins this ratio increases

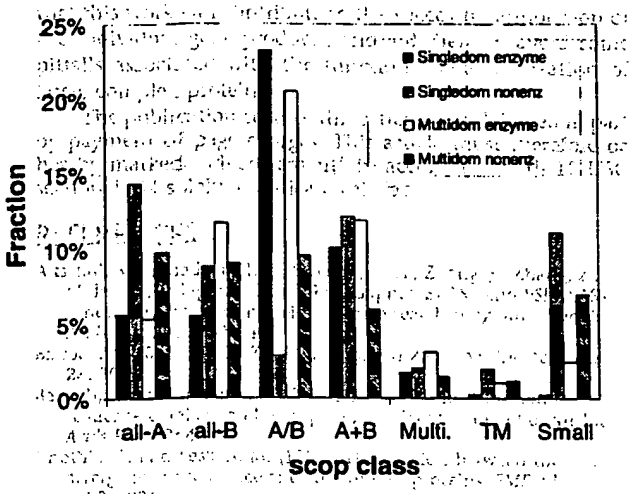


**Figure 3** Distribution of proteins amongst broad structural and functional classes; the distribution of the matches among the seven structural and two functional classes in single- and multi-domain proteins. The single-domain and multi-domain matches each total 100%, independently of each other. The horizontal axis indicates the seven scop classes, which are (from 1 to 7): all-alpha, all-beta, alpha/beta, alpha + beta, multi-domain, membrane, and small protein.

to almost threefold (135 vs. 56). This agrees with the notion that most enzymes are multi-domain. Another difference between single and multi-domain proteins appears in the ratio of superfamilies with a single function compared to multi-functional ones. As it is apparent from Table 1, about a quarter of the superfamilies matched single-domain proteins with different functions (55 of 215), whereas in the multi-domain proteins, this ratio increased to more than a third (119 of 310).

## Single-Domain Proteins

Table 2 lists the two functionally most diverse structural superfamilies in single-domain proteins with some representative functions. The most diverse superfamily, the 3.38.1 Thioredoxin-like, has 11 different functions associated with it, most of them with an oxidoreductase mechanism. For instance, THIO_BPT4 is a small disulphide-containing thioredoxin that serves as a general disulphide oxidoreductase, while TDX2_BRUMA is almost twice as long (199 aa) and serves as a thiol-specific antioxidant that acts against sulfur-containing radicals. Another interesting example of functional diversity is provided by the Scorpion toxin-like superfamily (7.3.6). While BRAZ_PENBA is a small protein that is known to be 2000 times sweeter than sucrose, the other members of the superfamily are associated with different host-defense mechanisms. In insects the superfamily possesses antifungal activity (DMYC_DROME) or acts as a toxin (SCX5_BUTEU). Interestingly, in plants it can also act as an antifungal (AF2B_SINAL) or as an inhibitor of insect alpha-amylases (SIA1_SORBI). It appears that many single-domain proteins are toxins or allergens, or are related in other ways to a host-defense response.

Based on the data we can also determine the probability of two single-domain proteins that match domains in the same superfamily category also carrying out the same function. Using Bayes' theorem:

$$P(F|S) = P(F)P(S|F)/(P(F)P(S|F) + P(\neg F)P(S|\neg F)) \quad (1)$$

where $S$ is the probability that two proteins share the same superfamily, $F$ is the probability that two proteins have the same function, and $\neg F$ is the probability that two proteins do not have the same function. Rearranging and simplifying the equation we get:

$$P(F|S) = 1/(1 + N(S,\neg F)/(N(S,F)) \quad (2)$$

where N is the number of times that the two events in the parentheses occur together in our database of 1818 single-domain proteins. This results in

$$P(F|S) = 1/(1 + 8501/12516) = 68\%.$$

That is, the probability that two single-domain proteins that have the same superfamily structure have the same function (whether enzymatic or not) is about 2/3.

## Multi-Domain Proteins

Table 3 lists the combinations of superfamilies that have been associated with the greatest number of different functions in multi-domain proteins, with representative entries in SWISS-PROT. The combination with the greatest number of different functions is that of 1.95.1 and 7.33.1. Although it has twice as many different functions as the most diverse superfamily in

**Table 1. Functional Distribution of Single-domain, Multi-domain Superfamilies, and Multi-domain Combinations**

| | Single-domain superfamilies | | Multi-domain superfamilies | | Multi-domain sfam combinations | |
|---|---|---|---|---|---|---|
| | Single function | Multiple function | Single function | Multiple function | Single function | Multiple function |
| Enzymatic | 82 | 11 | 135 | 42 | 151 | 16 |
| Nonenzymatic | 78 | 23 | 56 | 30 | 70 | 27 |
| Both functions | — | 15 | — | 47 | — | 17 |
| Total | 160 | 55 | 191 | 119 | 221 | 60 |

The basic functional distribution of the superfamilies in single- and multi-domain proteins and the functional distribution of multi-domain combinations are shown. The first row lists the number of scop superfamilies that were associated only with enzymatic function in each category. The second row lists the number associated with only nonenzymatic functions, and the third row indicates the number of superfamilies that were associated with both types of function. Altogether, we characterized 160 + 55 = 215 single-domain and 191 + 119 = 310 multi-domain superfamilies, 70 of which overlapped in the two categories.

**Table 2. Most Versatile Single-Domain Superfamilies**

| No. func | No. prot | Sfam comb | Function | SWISS-PROT ID | SWISS-PROT function |
|---|---|---|---|---|---|
| 11 | 69 | 3.38.1 | E1.11.1 | GSHP_RAT | Plasma Glutathione Peroxidase (1.11.1.9) |
| | | | 263# | DYL5_CHLRE | Dynein, Flagellar Outer Arm–*C. reinhardtii* |
| | | | D260# | BSAA_BACSU | Glutathione Peroxidase Homolog Bsaa |
| | | | 268# | REHY_TORRU | Rehydrin–*Tortula ruralis* (Moss) |
| | | | 266# | PHOS_HUMAN | Phosducin (33 Kd Phototransducing Protein) |
| | | | 269# | REHY_ORYSA | Rad24 Protein–*Oryza sativa* (Rice) |
| | | | 272# | THIO_BPT4 | Thioredoxin (Bacteriophage T4) |
| | | | D271#272# | TDX2_BRUMA | Thioredoxin Peroxidase 2 |
| | | | 261# | BTUE_ECOLI | Vitamin B12 Transport Periplasmic Protein Btue |
| 10 | 28 | 7.3.6 | 342# | BRAZ_PENBA | Brazzein–*Pentadiplandra brazzeana* |
| | | | 376#336# | SCKK_TITSE | Neurotoxin Ts-Kapa (Tsk)–(Brazilian scorpion) |
| | | | 341#356# | AF2B_SINAL | Cysteine-Rich Antifungal Protein 2b (Afp2b) |
| | | | 343# | DEFA_ZOPAT | Defensin, Isoforms B And C–*Zophobas atratus* |
| | | | 361# | DMYC_DROME | Drosomycin Precursor (Cysteine-Rich Peptide) |
| | | | 361#376# | SCX5_BUTEU | Insectotoxin I5a–(Lesser Asian scorpion) |
| | | | 336# | SCX3_LEIQH | Leiuropeptide Iii–(Scorpion) |
| | | | 203# | SIA1_SORBI | Small-Pr Inhibitor Of Insect Alpha-Amylases |
| 7 | 34 | 4.79.3 | 310# | AB18_PEA | Aba-Responsive Protein Abr18–Garden Pea |
| | | | 311# | DRR3_PEA | Disease Resistance Response Protein Pi49 |
| | | | 231# | MPAA_CORAV | Major Pollen Allergen Cor A 1,–Eu. Hazel |
| | | | 312# | L18B_LUPLU | Protein L1r18b (Lpr10.1b) |
| | | | E3.1.– | RNS2_PANGI | Ribonuclease 2 (3.1.–/–)–Panax Ginseng |
| | | | 314# | SAM2_SOYBN | Stress-Induced Protein Sam22 |
| 7 | 43 | 1.26.1 | 184# | CSF2_SHEEP | Colony-Stimulating Factor |
| | | | 381#564#184# | IL4_RAT | Interleukin-4 (B-Cell Igg Diff. Factor) |
| | | | 185# | LIF_HUMAN | Leukemia Inhibitory Factor (Lif) |
| | | | 187# | PRL_ANGAN | Prolactin Precursor (Prl)– |
| | | | 186# | PLF3_MOUSE | Proliferin 3 Mitogen-Regulated |
| | | | 188# | SOMA_PAROL | Somatotropin (Growth Hormone) |

The most versatile superfamilies in single-domain proteins as determined from their functional description in SWISS-PROT, with some representatives. The keyword combinations in the fourth column were based either on the first three components of their EC numbers (for enzymes) or derived automatically by comparing the DE description line of SWISS-PROT entries to a list of synonymous keywords at http://bioinfo.mbb.yale.edu/partslist/func. A keyword number starting with a D indicates an enzyme that does not have an assigned EC number in its description in SWISS-PROT.

the single-domain proteins (22 vs. 11, respectively), careful examination reveals that all the proteins in this category are DNA-binding and most of them act as hormone receptors.

The second entry listed in the table is the combination of the 3.4.1 and 4.48.1 superfamilies associated with the FAD/NAD(P)-linked reductases. It is an all-enzymatic combination and always carries out an oxido-reductase function. All the proteins in this category are completely covered by matches with these two superfamilies. The 1.78.1–2.1.1 hemocyanin-immunoglobulin combination seems also to be fairly conserved; although the proteins in this category are called by eight different names, most of them turn out to be extracellular larval storage proteins, except for the copper-containing oxygen carrier hemocyanin itself (HCY_PALVU).

Following the same logic, we can also determine the probability that two proteins that have the same superfamily combination share the same function, viz:

$$P(F|S) = 1/(1 + 32242/134230) = 81\%$$

This means that we have significantly greater certainty in determining the function of a multi-domain protein with a particular superfamily combination than that of a single-domain protein containing a particular superfamily. We also determined a similar probability for those proteins that have an almost complete coverage with exactly the same type and number of superfamilies, following each other in the same order. The probability that the functions are the same in this case was 91%, a considerably higher value than above. However, if two multi-domain proteins share only a single superfamily, the probability that they share the same function drops to only 35%! This greater functional certainty from sharing a combination of superfamilies rather than just one is also reflected in Table 1. While one-fourth of the single-domain proteins and one-third of singularly matching superfamilies in multi-domain proteins have multiple functions, only about one-fifth of the multi-domain *combinations* possess multiple functions (60 of 281). It is also clear from the data that domains in larger proteins often lose their original function and no longer have an autonomous function.

## Seventy Common Superfamilies and Their Functions Compared in Single-Domain and Multi-Domain Proteins

As mentioned above, of the 455 superfamilies in our analysis, only 70 occur in both single- and multi-domain proteins. Even more surprising is the small number of structural superfamilies (14) that have the same function in both single- and

**Table 3. Most Versatile Superfamily Combinations in Multi-Domain Proteins**

| No. func. | No. prot. | Sfam comb. | Function | SWISS-PROT ID | SWISS-PROT function |
|---|---|---|---|---|---|
| 22 | 176 | 1.95.1/7.33.1 | 29# | THB_RANCA | Thyroid Hormone Receptor Beta |
| | | | 10# | HNF4_DROME | Transcription Factor HNF-4 Homolog |
| | | | 31#32# | EAR2_MOUSE | V-Erba Related Protein Ear-2 |
| | | | 29#30# | ECR_MANSE | Ecdysone Receptor (Ecdysteroid Receptor) |
| | | | 32# | ERBA_AVIER | Erba Oncogene Protein |
| | | | 556#564#35# | NGFI_XENLA | Nerve Growth Factor Induced Protein I-B |
| | | | 576# | NR42_HUMAN | Immediate-Early Response Protein Not |
| | | | 36# | PPAT_HUMAN | Peroxisome Proliferator Activated Receptor |
| | | | 37# | RXTG_CHICK | Retinoic Acid Receptor RXR-Gamma |
| | | | 38# | TLL_DROVI | Tailless Protein |
| 8 | 54 | 3.4.1/4.48.1 | E1.8.2 | DHSU_CHRVI | Sulfide Dehydrogenase (1.8.2.–) |
| | | | E1.8.1 | DLDH_ZYMMO | Dihydrolipoamide Dehydrogenase (1.8.1.4) |
| | | | E1.6.4 | TYTR_TRYCR | Trypanothione Reductase (1.6.4.8) (Tr) |
| | | | E1.16.1 | MERA_STRLI | Mercuric Reductase (1.16.1.1) |
| | | | E1.6.99 | NAOX_MYCPN | Probable NADH Oxidase (1.6.99.3) (Noxase) |
| 8 | 23 | 1.78.1/2.1.1 | 19# | ARYB_MANSE | Arylphorin Beta Subunit–(Tobacco Hornworm) |
| | | | 20# | CRPI_PERAM | Allergen Cr-Pi Precursor–(American Cockroach) |
| | | | 21#427# | HCY_PALVU | Hemocyanin–(European Spiny Lobster) |
| | | | 22# | HEXA_BLADI | Hexamerin Precursor–(Tropical Cockroach) |
| | | | 23# | JSP1_TRINI | Acidic Juvenile Hormonne–Suppressible Protein |
| | | | 24# | LSP2_DROME | Larval Serum Protein 2 Precursor (LSP-2) |
| | | | 546#25# | SSP1_BOMMO | Sex-Specific Storage-Protein 1 |

Note that the combination with the greatest number of different functions is that of 1.95.1 and 7.33.1. Careful examination reveals that all the proteins with this combination are DNA-binding and most of them act as various hormone receptors. In particular, HNF4_DROME and NR42_HUMAN also have transcription activator functions. Note that these two proteins are considerably longer than the others in this group and are not covered completely by structural matches: A large C-terminal and a large N-terminal portion are left uncovered, respectively.

multi-domain proteins. These are listed in Table 4; 12 of them have enzymatic function, supporting the notion that enzymes are more conserved during evolution than non-enzymes. The two non-enzymatic superfamilies are the 4.29.1 ribosomal superfamily and the 5.4.1 superfamily in penicillin-binding proteins.

Table 5 presents several examples of the converse situation, shared superfamilies that have different functions in single and multi-domain proteins. Comparing parts A and B of the table highlights the fact that although both superfamilies in a multi-domain protein are often present in single-domain form as well, the functions in the different settings are only vaguely related. One example is the combination of the lipocalin superfamily (2.45.1) with that of the BPTI-like or Kunitz inhibitor (7.7.1), which in higher organisms forms a complex protein called alpha-1-microglobulin (AMBP_RAT). Another interesting example is the combination of the 2.5.1 Cupredoxin (occurring in the single-domain blue-copper protein, SOXE_SULAC) and the 6.5.1 Membrane all-alpha (single-domain representative: BACT_HALVA, a sensory rho-

**Table 4. Superfamilies With the Same Function in Single- and Multi-Domain Proteins as Determined from Their Keyword Combination or First Three Components of Their EC Numbers**

| | | Single-domain proteins | | Multi-domain proteins | |
|---|---|---|---|---|---|
| Sfam | Function | SWISS-PROT ID | SWISS-PROT function | SWISS-PROT ID | SWISS-PROT function |
| 1.81.1 | E3.2.1 | GUNY_ERWCH | Endoglucanase (3.2.1.4) | AMYG_NEUCR | Glucoamylase Precursor (3.2.1.3) |
| 2.66.2 | E3.5.1 | URE2_YERPS | Urease Beta (3.5.1.5) | URE1_HELPY | Urease Alpha Subunit (3.5.1.5) |
| 3.17.2 | E6.3.5 | NADE_MYCPN | NAD(+) Synthetase (6.3.5.1) | GUAA_YEAST | GMP Synthase (6.3.5.2) |
| 3.37.1 | E3.1.3 | PTP2_NPVOP | Protein-Tyrosine Phosphatase 2 (3.1.3.48) | PTNB_RAT | Protein-Tyrosine Phosphatase (3.1.3.48) |
| 3.67.1 | E4.2.1 | TRPB_VIBPA | Tryptophan Synthase (4.2.1.20) | TRP_YEAST | Tryptophan Synthase (4.2.1.20) |
| 4.19.1 | E5.2.1 | FKB1_METJA | Peptidylprolyl *Cis-Trans* Isomerase (5.2.1.8) | FKB7_WHEAT | 70 Kd Peptidylprolyl Isomerase (5.2.1.8) |
| 4.2.1 | E3.2.1 | LYCV_BPP2 | Lysozyme (3.2.1.17) | CHIX_PEA | Endochitinase Precursor (3.2.1.14) |
| 4.29.1 | 85# | RS5_ACYKS | 30s Ribosomal Protein S5 | RS5_TREPA | 30s Ribosomal Protein S5 |
| 4.52.1 | E3.4.24 | SNPA_STRCS | Extracellular Neutral Protease (3.4.24.–) | BMPH_STRPU | Collagenase 3 Precursor (3.4.24.–) |
| 4.6.1 | E3.5.1 | URE3_YERPS | Urease Gamma (3.5.1.5) | URE1_HELPY | Urease Alpha Subunit (3.5.1.5) |
| 5.10.1 | E2.7.7 | KANU_STAAU | Kanamycin Nucleotidyltransferase (2.7.7.–) | DPOB_XENLA | Dna Polymerase Beta (2.7.7.7) |
| 5.4.1 | 161# | AMPH_ECOLI | Penicillin-binding Protein Amph | PBPX_STRPN | Penicillin-binding Protein 3x Pbp2x |

**Table 5. Examples of Superfamilies Present In Both Single- and Multi-Domain Proteins, Carrying out Different Functions**

**Table 5A. Single-Domain Proteins**

| Sfam | Funct # | SWISS-PROT ID | SWISS-PROT function |
|---|---|---|---|
| 1.25.1 | 352# | FTN2_HAEIN | Ferritin-like Protein 2 |
| | 183# | NIGY_DESVH | Nigerythrin |
| | E1.17.4 | RIR4_YEAST | (Ribonucleotide Reductase) (1.17.4.1) |
| | 192# | NLP_HAEIN | Ner-like Protein Homolog |
| 1.4.3 | 196# | H1A_PLADU | Histone H1A, Sperm |
| 1.81.2 | E2.5.1 | PFTB_PEA | Farnesyltransferase Beta Su (2.5.1.–) |
| 2.45.1 | 226# | ERBP_RAT | Epididymal-Tetinoic Acid Binding Protein |
| | 227# | FAB3_CAEEL | Fatty Acid-Binding Protein Homolog 3 |
| | 228#412# | NGAL_MOUSE | Neutrophil Gelatinase-Assoc. Lipocalin |
| | 229# | NP4_RHOPR | Nitrophorin 4 Precursor |
| | E5.3.99 | PGHD_HUMAN | Prostaglandin-H2 D-Isomerase (5.3.99.2) |
| | 230#421# | VNS1_MOUSE | Vesomeral Secretory Protein I |
| 2.5.1 | 231# | MPA3_AMBEL | Pollen Allergen AMB A 3 (AMB A III) |
| | 232#427# | SOXE_SULAC | Sulfocyanin (Blue Copper Protein) |
| 3.14.2 | 373# | RRF1_DESVH | Rrf1 Protein |
| 3.29.1 | E6.3.4 | PURA_CAEEL | Adenylosuccinate Synthetase (6.3.4.4) |
| | E2.7.4 | KTHY_YEAST | Thymidylate Kinase (2.7.4.9) |
| | D259# | VA57_VACCV | Guanylate Kinase Homolog |
| | E2.7.1 | KITH_VZVW | Thymidine Kinase (2.7.1.21) |
| 3.47.1 | 275# | MBL_BACSU | MBL Protein |
| | 276# | MREB_BACSU | Rod Shape-determining Protein Mreb |
| 3.48.1 | E3.1.3 | PPA5_YEAST | Repressible Acid Phosphatase (3.1.3.2) |
| 3.81.1 | D281# | AMIC_PSEAE | Aliphatic Amidase Expression-Regulator |
| | 282# | LUXP_VIBHA | LUXP Protein Precursor |
| 4.103.1 | E2/4/2 | TOX1_BORPE | Pertussis Toxin Su 1 (2.4.2.–) |
| 4.105.1 | 291# | LECC_POLMI | Lectin–Polyandrocarpa Misakiensis |
| 4.11.5 | 295# | TERP_PSESP | Terpredoxin |
| 4.19.1 | E5.2.1 | FKB1_METJA | Pept-Prolyl *Cis-Trans* Isomerase (5.2.1.8) |
| 6.5.1 | E3.6.1 | ATPL_VIBAL | ATP Synthase (3.6.1.34) (Lipid-binding) |
| | 540#325# | BACT_HALVA | Sensory Rhodopsin II (Sr-II) |
| 7.35.4 | E1.9.3 | COXB_RAT | Cytochrome C Oxidase (1.9.3.1) (Via*) |
| | 345# | DESR_DESBI | Desulforedoxin (Dx) |
| 7.7.1 | 349# | TAP_ORNMO | Tick Anticoagulant Peptide |

*(Table continues on following page.)*

dopsin) superfamilies into a component of the respiratory chain, cytochrome C oxidase II (COOX_ZOOAN). All these examples demonstrate the evolutionary advantage of a domain fusion event, which creates a function that is more complex than either of the components.

## Multifunctionality vs. Sequence Similarity

Previously, we presented a variety of graphs that show how the probability that two domains would share the same function varied with respect to sequence similarity (Hegyi and Gerstein 1999; Wilson et al. 2000). Figure 4 shows a similar graph with the calculations extended to multi-domain proteins. The figure shows that the functional divergence of a single domain in multi-domain proteins dramatically increases, more than twofold, compared to the single-domain ones. This reinforces our findings above, based only on superfamily content, that the certainty with which we can predict the function of a protein based on its sequence similarity with a domain in another multi-domain protein, is considerably less than for a comparable single-domain situation.

**Table 5B.** Multi-Domain Proteins

| Sfam Comb. | Funct# | SWISS-PROT ID | SWISS-PROT function |
|---|---|---|---|
| 1.25.1/7.35.4 | 104# | RUBY_METJA | Putative Rubrerythrin |
| 1.32.1/3.81.1 | 11# | PURR_HAEIN | Purine Nucleotide Synthesis Repressor |
| | 12# | DEGA_BACSU | Degradation Activator |
| | 581#11# | SCRR_STRMU | Sucrose Operon Repressor |
| | 582#11# | REGA_CLOAB | Transcription Regulatory Protein Rega |
| 1.4.3/3.14.2 | 10# | SKN7_YEAST | Transcription Factor Skn7 (Pos9 Protein) |
| | 11# | VIRG_AGRT5 | Virg Regulatory Protein |
| | 13# | RGX3_MYCTU | Sensory Transduction Protein REGX3 |
| | 190# | PFER_PSEAE | Transcriptional Activator Protein Pfer |
| | 366# | PETR_RHOCA | Petr Protein |
| 2.45.1/7.7.1 | 203#153# | HC_RAT | Alpha-1-Microglobulin/Trypsin Inhibitor |
| 2.5.1/6.5.1 | E1.9.3 | COX2_ZOOAN | Cytochrome C Oxidase II (1.9.3.1) |
| 3.29.1/3.48.1 | E2.7.1 | F26_RANCA | 6-Phosphofructo-2-Kinase (2.7.1.105) |
| 3.47.1/5.17.1 | 1# | YED0_YEAST | Heat Shock Protein 70 Homolog YEL030w |
| | 1#83# | GR73_MAIZE | Ig-Binding Protein |

## DISCUSSION

Here we built on our previous studies on the relationship between protein structure and function to develop new results related to multi-domain proteins. Throughout the paper, we focused on superfamilies instead of folds, as the members of a superfamily are presumably of common evolutionary origin (Murzin et al. 1995).

We found that the 4763 multi-domain and 1818 single-domain proteins that met our selection criteria have about the same distribution of structural classes, with more enzymatic functions associated with the alpha/beta structural classes and more non-enzymatic ones with the all-alpha and small classes. We identified more than three times as many multi-domain proteins that were enzymes than single-domain ones (2805 and 850, respectively) and, conversely, about twice as many multi-domain proteins as single-domain ones that were non-enzymes (1958 vs. 968).

We focused on the functional divergence of the two groups and found that about a quarter of the superfamilies in single-domain proteins are associated with multiple functions, whereas only about a fifth of the multi-domain superfamily combinations are. Therefore, we can conclude that a combination of specific superfamilies results in a more specific functional assignment for a particular protein. However, about one-third of the superfamilies in the multi-domain proteins were associated with multiple functions, underlining the lesser autonomy of a domain function in multi-domain protein.

This latter finding was also supported by the difference in functional divergences between the two groups of proteins based on particular sequence similarities between the domains and SWISS-PROT proteins. As is shown in Figure 4, the average functional divergence of a single domain is much larger (more than twofold) in multi-domain proteins than in single-domain ones.

We also found that only 70 of a total of 455 superfamilies are shared between the multi-domain and single-domain proteins and only a small fraction (14) share their functions. This was rather surprising to us, and should be taken into consideration in functional characterization and annotation of new gene products. When the functions were related in single- and multi-domain proteins, we could observe an increasing functional complexity with the appearance of large multi-domain proteins.

Altogether, with the recent sequencing of the human genome and the genomes of other model organisms, we hope

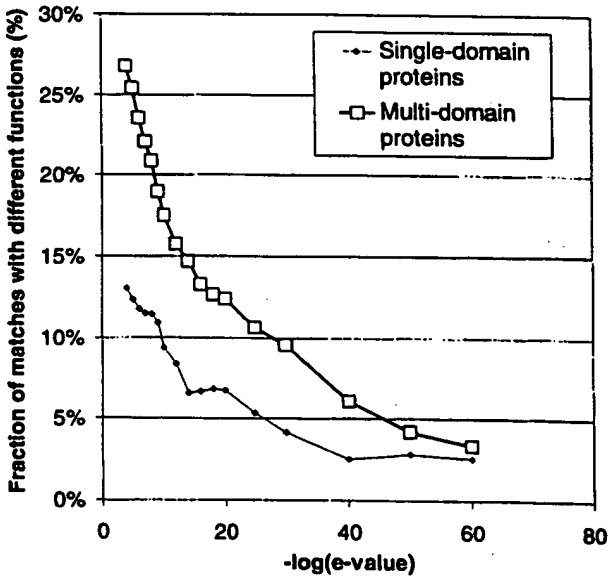


**Figure 4** Divergence in function with respect to sequence similarity. Relative number of matching domains with multiple functions, as the function of *e*-value threshold. Diamonds represent single-domain proteins, squares multi-domain ones (matching just for a single domain), respectively. The first value on the *X*-axis starts at 4 (corresponding to an *e*-value=$10^{-4}$).

that this work can contribute to the successful annotation of the individual gene products, and will help to avoid some pitfalls associated with the functional characterization of large, complex proteins.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

## REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Bairoch, A. 2000. The ENZYME database in 2000. *Nucleic Acids Res.* **28:** 304–5.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–8.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5:** 823–826.

Devos, D. and Valencia, A. 2000. Practical limits of function prediction. *Proteins* **41:** 98–107.

Drawid, A. and Gerstein, M. 2000. A Bayesian system integrating expression data with sequence patterns for localizing proteins: Comprehensive application to the yeast genome. *J. Mol. Biol.* **301:** 1059–1075.

Eisenstein, E., Gilliland, G. L., Herzberg, O., Moult, J., Orban, J., Poljak, R. J., Banerjei, L., Richardson, D. and Howard, A. J. 2000. Biological function made crystal clear - annotation of hypothetical proteins via structural genomics. *Curr. Opin. Biotechnol.* **11:** 25–30.

Gelbart, W. M., Crosby, M., Matthews, B., Rindone, W. P., Chillemi, J., Russo Twombly, S., Emmert, D., Ashburner, M., Drysdale, R. A., et al. 1997. FlyBase: A *Drosophila* database. The FlyBase consortium. *Nucleic Acids Res.* **25:** 63–6.

Gerstein, M. 1997. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274:** 562–76.

———. 1998. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des.* **3:** 497–512.

Harrison, P. Echols, N. and Gerstein, M. 2001. Digging for dead genes: An analysis of the characteristics of the pseudogene population in the *C. elegans* genome. *Nucleic Acids Res.* **29:** 818–830.

Hegyi, H. and Gerstein, M. 1999. The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288:** 147–164.

Lin, J. and Gerstein, M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels. *Genome Res.* **10:** 808–818.

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. and Thornton, J. M. 1998. Protein folds and functions. *Structure* **6:** 875–884.

Murzin, A., Brenner, S. E., Hubbard, T. and Chothia, C. 1995. SCOP: A structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Orengo, C. A., Pearl, F. M., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L. and Thornton, J. M. 1999. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.* **27:** 275–279.

Pawlowski, K., Jaroszewski, L., Rychlewski, L. and Godzik, A. 2000. Sensitive sequence comparison as protein function predictor. *Pac. Symp. Biocomput.* 42–53.

Pearson, W. R. 1994. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.* **25:** 365–389.

Qian, J., Stenger, B., Wilson, C., Lin, J., Jansen, R., Krebs, W., Alexandrov, V., Echols, N., Teichmann, S., Park, J. et al.. 2001. PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res.* **29:** 1750–1764.

Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A. and Sternberg, M. J. 1997. Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *J. Mol. Biol.* **269:** 423–439.

Shah, I. and Hunter, L. 1997. Predicting enzyme function from sequence: A systematic appraisal. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5:** 276–283.

Shapiro, L. and Harris, T. 2000. Finding function through structural genomics. *Curr. Opin. Biotechnol.* **11:** 31–5.

Stawiski, E.W., Baucom A.E., Lohr S.C., and Gregoret L.M. 2000. Predicting protein function from structure: Unique structural features of proteases. *Proc. Natl. Acad. Sci.* **97:** 3954–8.

Thornton, J. M., Orengo, C. A., Todd, A. E. and Pearl, F. M. 1999. Protein folds, functions and evolution. *J. Mol. Biol.* **293:** 333–342.

Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. and Orengo, C. A. 2000. From structure to function: Approaches and limitations. *Nat. Struct. Biol.* **7 Suppl:** 991–994.

Todd A.E., Orengo C.A., and Thornton J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307:** 1113–1143.

Wilson, C. A., Kreychman, J. and Gerstein, M. 2000. Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297:** 233–249.

Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J. S., Skolnick, J. and Godzik, A. 1999. From fold predictions to function predictions: Automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8:** 1104–1115.

*Received February 7, 2001; accepted in revised form June 19, 2001.*